Universidad Autónoma de Madrid

Departamento de Biología Molecular

Tesis Doctoral

Online Assessment of Protein Interaction Information Extraction Systems

Memoria presentada para optar al grado de Doctor en Ciencias por:
Florian Leitner

Director de Tesis:
Prof. Alfonso Valencia

Madrid, 2011

Acknowledgements

I wish to thank my scientific guide and director of this thesis, Prof. Alfonso Valencia, for his support in the development of my scientific career. He has been a most excellent supervisor during these years, I could have wished for nothing more. His holistic understanding of molecular biology has provided my work with its scientific foundations and rigor.

Much of this work would not have been possible without all the scientific discourse I shared with my colleague Martin Krallinger. His comprehensive understanding of the field of natural language processing in molecular biology proved essential for this work. In addition, I wish to acknowledge all members of Prof. Valencia's research group for the interesting environment and all the helpful advice they have provided. In particular, I would like to name the following persons for contributions to parts of this work: Miguel Vázquez, José-Manuel Rodriguez, Eduardo León, and Angel Carro.

All BioCreative co-organizers and participants as well as the database curators we collaborated with should be acknowledged here, too. None of this work would have been possible without this great scientific community.

Finally, I wish to thank my family – especially my late mother Julie A. Stone-Leitner and my father Harald H. Leitner – for all their love and support. And my last – but not least – thoughts go to my finacée, Mayte A. Valdez-Roitenburd, for her continuous encouragement and her enduring patience during the months of writing this thesis.

Resumen

Antecedentes. La extracción manual de interacciones entre proteínas (PPIs) que realizan los curadores de las bases de datos (BBDD) no puede abarcar toda la literatura científica sobre dichas interacciones. Tampoco es previsible que la proteómica de alto rendimiento pueda reproducir estos datos en un futuro próximo. Parte de la información presente en estas BBDD de PPI podría extraerse automáticamente de la literatura utilizando herramientas de minería de textos. A través de las últimas tres competiciones BioCreative (II, II.5 y III) se han podido realizar evaluaciones colectivas de estos sistemas de minería de textos de PPI. Estas competiciones ("community challenges", ver Tabla 1) contribuyen al intercambio de conocimiento científico, facilitan la definición de los estándares de anotación, así como la creación de *corpora* duraderos, útiles para el desarrollo de métodos de minería de texto. Al mismo tiempo, el experimento de FEBS Letters, concerniente a los resúmenes digitales estructurados (SDAs), ha servido para evaluar la utilidad de incluir a los autores en el proceso de curación. En conjunto, los resultados representan la base para comparar la extracción de datos de PPI a partir de la literatura, ya sea mediante minería de textos, autores o curadores de BBDD.

Resultados. Los temas abarcados en las competiciones incluían la clasificación de artículos de PPI, la asignación de identificadores de BBDD a las menciones de proteínas, la detección de interacciones y la extracción de métodos experimentales. En total, BioCreative II, II.5 y III atrajo a 52 grupos de investigación de todo el mundo. El BioCreative Meta-Server (BCMS) recopilaba las anotaciones de los participantes de BioCreative II y fue modificado para las dos competiciones siguientes con el fin de garantizar la participación online. Tres de las BBDD de IMEx (MINT, IntAct, y BioGRID) han producido anotaciones "gold standard" para un total de 3.632 artículos de texto completo y proporcionaron 19.642 clasificaciones de resúmenes como PPI-relevante o no. Introducimos una nueva métrica de evaluación – FAP_β – para puntuar listas ordenadas, compuesto por la media armónica entre "F_β-score" y "Average Precision" (AP). Penaliza a los conjuntos de resultados desproporcionados, y no requiere de un umbral. La evaluación mostró que la clasificación automatizada de artículos PPI podría reducir el tiempo necesario para seleccionar los artículos relevantes por un tercio. Los mejores sistemas de asignación lograron un F_1-score de 29% y un AP de 26%.

Conclusiones. Los "gold standard" *corpora* de BioCreative forman recursos duraderos, documentado por los cientos de participantes y citaciones registrados. El ordenamiento automático de artículos puede mejorar la selección de artículos PPI en comparación con una búsqueda basada en palabras clave. La causa principal de errores en la asignación de identificadores a menciones de proteínas se debe a fallos en la desambiguación de organismos. El mejoramiento de sistemas que detectan los correspondientes identificadores a las menciones de proteínas es necesario para lograr mejores resultados en la extracción de interacciones de pares de proteínas (normalizadas). Un número relevante de anotaciones producidas por curadores se basan en relaciones de discurso a distancia que no se abarcan por los enfoques actuales. Para mejorar la extracción de conceptos relacionados con métodos experimentales, hay que mapear las expresiones científicas utilizadas a los términos ontológicos. El FAP_β identifica los mejores resultados automáticos y puede establecer valores de corte para los enfoques con alta recuperación y baja precisión. Aunque los sistemas automatizados no generan anotaciones que coinciden con las de los de autores o curadores, proporcionar resultados consenso con un meta-servicio y una mínima intervención humana abre nuevas vías para ayudar a los curadores y autores en la recuperación de identificadores relevantes.

Abstract

Background. Manual database (DB) curation efforts extracting protein-protein interactions (PPIs) from publications are not able to cover the entire scientific literature on those interactions. Neither is high-throughput proteomics likely to reproduce this data in the foreseeable future. Existing text mining systems can extract a subset of the PPI data curated by DBs from publications. With the last three BioCreative challenges (II, II.5, and III), community evaluations of these PPI text mining systems have been made. Community challenges contribute to scientific exchange, define annotation standards, and generate lasting corpora to develop text mining approaches. At the same time, the FEBS Letters experiment on Structured Digital Abstracts (SDAs) evaluated the utility of incorporating authors in the curation process. Together, the joint results provide the basis for comparing the extraction of PPI data from literature using text mining, authors, or curators.

Results. Some tasks of the challenges were PPI article selection, mapping interacting proteins to DB identifiers, detecting interacting protein pairs, as well as reporting the experimental methods used by the authors. In total, BioCreative II, II.5, and III attracted 52 research groups world-wide. The BioCreative Meta-Server (BCMS) collated the annotations from BioCreative II participants and was modified for the two following challenges to hold the challenges online. The MINT, IntAct, and BioGRID IMEx databases curated the gold standard annotations for 3,632 full-text articles and provided 19,642 classifications of abstracts as PPI-relevant or not. We introduce a new evaluation function - FAP_β - for scoring ranked lists, composed of the harmonic mean between F_β-score and Average Precision (AP). It penalizes disproportionate result sets, and does not require a threshold. The evaluation showed that automated ranking of PPI articles could reduce the time needed to select relevant articles by one third. The best mapping systems achieved a F_1-score of 29% and an AP of 26%.

Conclusions. The BioCreative gold standard corpora form long-lasting assets, documented by hundreds of registered participants and citations. Article ranking can facilitate PPI article selection when compared to keyword-based queries. Better mapping systems are needed to improve the extraction quality of (normalized) interaction pairs, with the main mapping error arising from organism disambiguation issues. However, a relevant number of curator annotations are based on long-range discourse relations that are not captured by current approaches. Improving experimental method concept extraction will require mapping the scientific jargon to their ontology terms. The FAP-score identifies the best submissions and can establish cutoffs for high-recall/low-precision approaches. While automated systems did not generate annotations that match those of authors or curators, providing ensemble results from a meta-service and minimal human intervention point to new avenues for assisting curators and authors in retrieving relevant identifiers.

List of Abbreviations

AP Average Precision
AUC Area Under Curve
BCMS BioCreative Meta-Server
BioNLP Natural Language Processing in Biology
CRF Conditional Random Field
FAP F-measured Average Precision
FEBS Federation of European Biochemical Societies
FN False Negative
FP False Positive
HMM Hidden Markov Model
IAA Inter-Annotator Agreement
IE Information Extraction
IR Information Retrieval
JPA Joint Probability of Annotation
GN Gene Normalization
GO Gene Ontology
KGT Knowledge Generation from Text
Max. Ent. Maximum Entropy (i.e., multinomial logistic regression)
MCC Matthew's Correlation Coefficient
MeSH Medical Subject Headings (thesaurus)
MEMM Maximum Entropy Markov Model
NER Named Entity Recognition
NLM National Library of Medicine
NLP Natural Language Processing
PoS Part-of-Speech
PSI-MI Proteomics Standards Initiative - Molecular Interactions (ontology)
PPI Protein-Protein Interaction
PR Precision/Recall
ROC Receiver Operator Characteristic
SDA Structure Digital Abstract
SVM Support Vector Machine
TAP Threshold Average Precision
TF-IDF Term Frequency times Inverted Document Frequency
TN True Negative
TP True Positive
UMLS Unified Medical Language System

Contents

CHAPTER 1

Introduction

1.1 Natural Language Processing in Biology

1.1.1 Incentive

Many areas of molecular biology - commonly summarized by the "omics" neologism (Evans 2000, Ramachandran & Dash 1999) - have undergone significant transformations during the past two decades. Particularly, high-throughput methods have induced a large increase in experimental data since the turn of the millennium. In proteomics, the research on chip technology has introduced protein arrays (MacBeath & Schreiber 2000). Together with the early advances on two- and one-hybrid screening (Fields & Song 1989, Vidal et al. 1996), these methods are used for analyzing enzyme–substrate, DNA–protein and protein-protein interactions (Templin et al. 2002, Zhu & Snyder 2003). The first comprehensive protein–protein interaction (PPI) network generated with these interaction detection methods was the yeast interactome (Ito et al. 2001, von Mering et al. 2002). Metabolomics is exploring the interaction of the genome and proteome with the chemical environment of the cell (Fiehn 2002, Kell 2004) using mass spectrometry (MS) and nuclear magnetic resonance (NMR). Other technologic advances - such as sequencing the human genome (Venter et al. 2001) using Sanger sequencing and genome-wide microarray expression profiling (Brown & Botstein 1999) - have triggered similar transformations in genomics. Recently, sequencing technology has moved on to what is commonly referred to as the "next-generation sequencing" (Metzker 2010). All these large-scale experiments produce data that need to be maintained in a broad spectrum of biological repositories. To quote a number of existing repositories, the 2010 NAR database issue records 1230 different resources alone (Cochrane & Galperin 2010).

These emerging omics marked the beginning of personalized medicine (Bell 2004), where these hybridization, spectrometric, and electrophoretic methods are applied to individual patients samples (Hood et al. 2004). Even genomic profiling, e.g., with ChIP-seq, is becoming more feasible as the costs per sample are dropping (Park 2009, Drmanac 2011). Data from these patient assays are screened against known genomic, proteomic, metabolomic, and pharmacogenetic information (Khoury et al. 2003). The goal is detecting abnormalities that can aid in early disease diagnosis and prevention, or markers that can indicate beneficial or adverse drug effects (Weston & Hood 2004).

Experimental data are then integrated in the context of computational biology, extracting new insights and potential research targets (Joyce & Palsson 2006). Building on these "omics" advances, system biology began modeling molecular networks to describe

their emergent properties (Kitano 2002). Data mining approaches associate biological markers (e.g., short nucleotide polymorphisms (Mooney 2005)) or interactions (e.g., PPIs (Oti et al. 2006)) to disease phenotypes. Genetic variation and evolutionary conservation is explored to gain functional understanding of genes and pathways (White 2001, Erdman et al. 2006, ENCODE Project Consortium et al. 2007). In structural genomics, proteins are modeled in silico to understand their function (Baker 2001). With the advent of high-volume data, comprehensive data integration of large-scale networks yield new functional insights. A very recent example is the work of (Kiel et al. 2011), uncovering the cellular response network of the rhodopsin G protein-coupled receptor.

For all these approaches, data needs to be available in structured repositories that index and maintain the content in a machine-readable format. Despite the large number of database projects cataloging research results, they are not at par with the information provided by scientific publications. For example, BioGRID currently catalogs nearly 49,000 human protein interactions taken from slightly over 10,000 publications (Stark et al. 2011). Estimates for the size of the human interactome predict about 650,000 interactions (Stumpf et al. 2008) or as low as 200,000 (Peri et al. 2003). The number of reported interactions in literature will fall between the estimates and the interactions already recorded in the PPI databases. Furthermore, the number of interactions in the protein interaction databases is far lower for most other organisms than human. Because of the bottleneck of transforming written publications into relational data schemata, the majority of all generated biomedical information remains exclusive to the publications (Chatr-aryamontri et al. 2008).

This provides the fundamental incentive of applying text mining methods to biomedical articles – to extract structured data from the biomedical literature with the goal of reducing this disparity. Text mining has been helping biologists trace genetic interactions (Hoffmann et al. 2005) and disease-relevant mutations (Krallinger et al. 2009), extracting phenotype-genotype relationships (Perez-Iratxeta et al. 2005) and entire interaction networks from literature (Blaschke & Valencia 2001), or aiding in the discovery of new hypotheses (Tsuruoka et al. 2008).

In other words, a substantial amount of biomedical data escapes analysis because they are contained exclusively in the scientific literature (Martin-Sanchez et al. 2004). For example, to populate PubMeth, a cancer methylation database (Ongenaert et al. 2008), text mining methods are being applied to speed up the selection of abstracts and to retrieve relevant database identifiers. Pharmacogenomic, -dynamic and -kinetic

information used to assess drug toxicity is extracted with information extraction methods (Rubin et al. 2005, Wang, Kim, Quinney, Guo, Hall, Rocha & Li 2009). Text mining is assisting in areas such as drug discovery (Loging et al. 2007) and is used to extract drug-drug interactions (Segura-Bedmar et al. 2011*b*). It has been applied to patient records, extracting associations between diseases and genetic disorders (Lage et al. 2007). While the data generated from molecular profiling creates the empirical foundation for diagnosis, doctors will require more than probabilities and risk factors to judge the possible avenues of treatment indicated by these results. Text mining will be the key to augment or support the known relationships, reporting contextual background knowledge for associations between known disease markers and the patient assay data (Krallinger, Leitner & Valencia 2010).

Another example of applied text mining that has become very popular over the last decade is the analysis of DNA microarray data (Tanabe et al. 1999). The data from differentially expressed genes need to be characterized to explain the corresponding phenotype. For this, available information about these genes has to be correlated with the observed data (Park et al. 2001). To explain the data, it is necessary to explore functional and pathway information about these genes, their tissue-specificity, protein localization and interaction data, transcription regulation information, co-expression data, and disease associations. However, little of this data is available in structured repositories, and text mining has been used to cluster and associate genes with the biomedical literature (Chaussabel & Sher 2002). To provide an investigator with necessary clues, text mining is used to supplement the profile data with this information, extracted from the scientific literature (Pedicini et al. 2010).

A similar situation exists in the context of RNA interference perturbation experiments. They are used for genome-scale, targeted knock-downs of genes to systematically explore their function (Kiefer et al. 2009). The effects are monitored at various endpoints, e.g., via measuring promoter activation or cell survival and proliferation. Again, relevant context information can often only be found in the literature (Sales et al. 2010). Extracting this data from literature is necessary to make hypotheses about the experimental observations.

Regarding the protein interaction data that has been collected over the past decades and reported in the literature, it might be argued that the advance of high-throughput methods (in proteomics) has enabled us to reproduce these interactions. This would indicate that the interaction relationships between proteins could be reconstructed using

high-throughput techniques such as two-hybrid screening or mass spectrometry (Lalonde et al. 2008). A recent high-throughput study supported by 46 scientists across 18 institutes attempted to extract a map of all interactions between transcription factors in mouse and human (Ravasi et al. 2010). In this study, the protein interactions between known transcription factors were measured with large-scale mammalian two hybrid (M2H) screens. Starting with a set of 1988 human and 1727 mouse factors, they were able to detect 762 and 877 interactions in human and mouse, respectively. Then, investigating 289 interactions recorded in public databases, the authors selected 91 interactions supported by at least two independent studies. However, only one quarter of those interactions had been detected during the M2H screening, and only half of 34 randomly chosen interactions detected by the M2H approach could also be confirmed by in vitro pull-down. This does not invalidate their approach, as these protein-protein interactions (PPIs) might be transient or unstable under the pull-down conditions. However, these large-scale screens have been criticized for producing large numbers of false positives (Sprinzak et al. 2003). Apart from precision, the recovery rate does show that the large majority of biological information about PPIs remains exclusive to the publications. Despite their power, high-throughput methods for now do not seem to fully reproduce this knowledge. Furthermore, the most reliable interaction data are generated by studies of individual interactions with high confidence methodologies such as X-ray crystallography, co-immunoprecipitation, or protein cross-linking (Miernyk & Thelen 2008) that are inherently small-scale. Therefore, recovering these relations from the literature is the only way to create a map of the interactome that fully reflects the current science.

In summary, text mining facilitates the access to biological knowledge (Cohen & Hersh 2005). Biomedical data integration approaches can benefit from text mining, retrieving and extracting information only available in the literature (Krallinger, Valencia & Hirschman 2008). Beyond making the information available for human interpretation, text mining assists in hypothesis generation and identifying new research targets (Jensen et al. 2006).

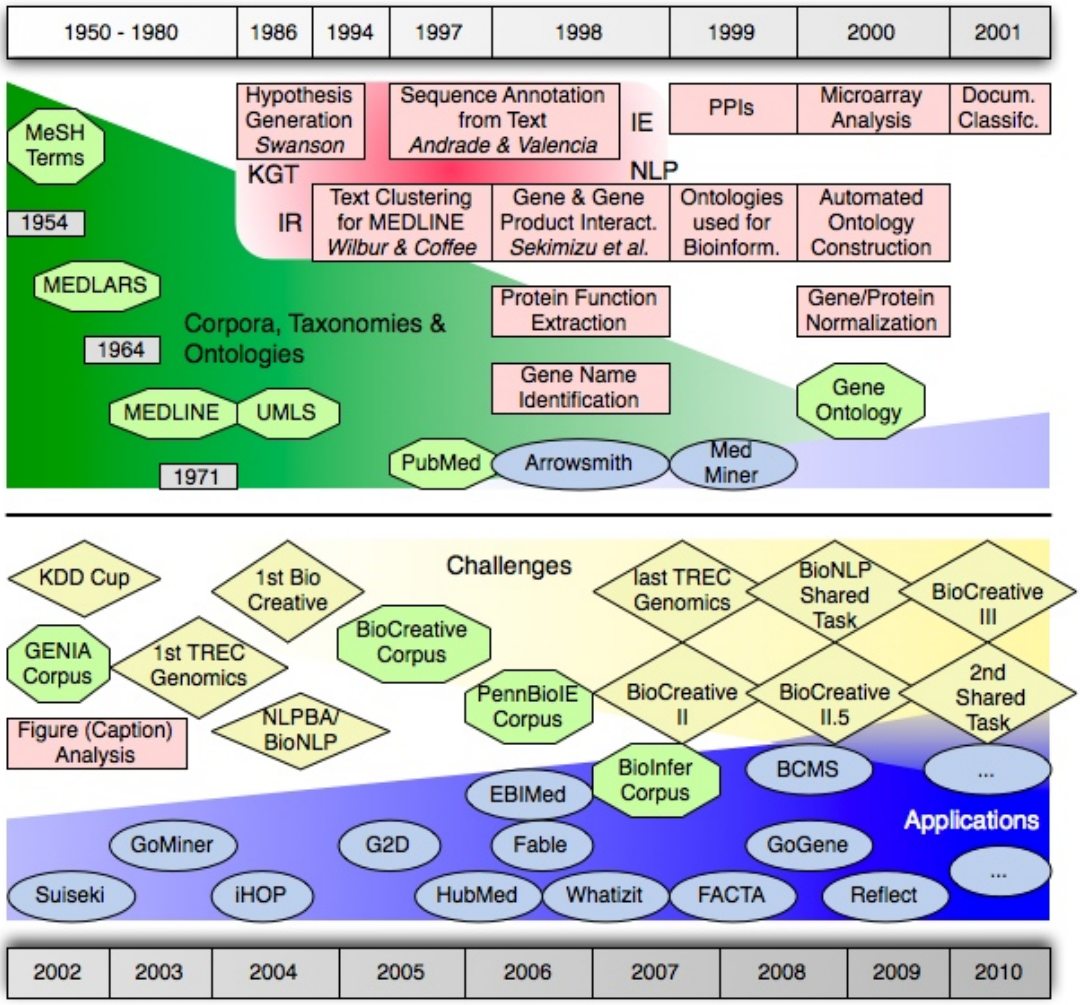

Figure 1.1: A timeline of the historic development of the BioNLP field. **Green:** development of MEDLINE/PubMed resources. **Red:** First publications of fundamental BioNLP methods. **Blue:** A selection of BioNLP applications. **Yellow:** Community challenges in BioNLP. See text for details. Abbreviations: Docum. Classific.: document classification; Bioinform.: bioinformatics

1.1.2 History and Scope

MEDLINE and PubMed

Most biomedical publications are recorded by the US National Library of Medicine (NLM) and made publicly available via PubMed[1]. Established in 1971 (as the MEDLINE database), by 2011, this collection of citations surpassed 21 million records pointing to biomedical literature[2]. Many PubMed records include the abstract of journal publications, in addition to meta-data such as author names, language, or journal information. A subset of the PubMed records provides bibliographic meta-data that is annotated semi-automatically by NLM curators for each citation. This meta-data, the Medical Subject Headings (MeSH) thesaurus (Rogers et al. 1963), was introduced even earlier, in 1954, to categorize the NLM literature and has been expanded into a detailed classification schema describing more than 2.5 million biomedical concepts in 21 different languages, known as the Unified Medical Language System (UMLS) (Bodenreider 2004). The PubMed collection of citation data, consisting primarily of publication titles and abstracts, author names, journal information, and MeSH categories, was used for text mining the first time in the late 1980s and early 90s and became instrumental in the development of the area of text mining and its linguistic aspect, **natural language processing in biology** (BioNLP). A timeline of events relevant to text mining in biology is shown in Figure 1.1. A short introduction to text mining and BioNLP terms can be found in the Appendix, Section A.3.

BioNLP Origins

The first publication using a text mining approach is commonly assigned to the subdiscipline of **knowledge generation from text** (KGT). KGT is centered on the concept of generating novel hypotheses using text mining. In 1986, Swanson was able to infer a relation between fish oil and Raynaud's syndrome using relationships mined from text (Swanson 1986): Raynaud patients have high blood viscosity and platelet aggregation, and they suffer from vasoconstriction. Fish oil, or rather its active ingredient eicosapentaenoic acid, lowers viscosity and aggregation, and causes vasodilation. Mining these two relationships from independent sources in the literature and using the vascular state as "pivot topic", Swanson demonstrated that via the pivot, it can be inferred that

[1] `http://www.ncbi.nlm.nih.gov/pubmed/`
[2] `http://www.nlm.nih.gov/bsd/revup/revup_pub.html`

fish oil aids patients suffering from Raynaud's disease. Swanson's approach later gave rise to the first text mining tool, Arrowsmith (Smalheiser et al. 2006).

The second main theme of text mining is **information retrieval** (IR). The techniques of retrieving biological text for a given query were introduced first in biology in 1994 by Wilbur and Coffee (Wilbur & Coffee 1994). They presented a method for clustering of the MEDLINE titles and abstracts by similarity to allow immediate retrieval of "neighboring" (i.e., similar) documents to the original query or currently viewed document.

The third sub-discipline in this general classification of text mining, **information extraction** (IE), made use of the MEDLINE titles and abstracts, too. In 1997, Andrade and Valencia designed a machine system to annotate the functional characteristics of protein families based on statistical text processing (Andrade & Valencia 1997). Relevant keywords for annotating a specific family are selected by comparing their frequency in abstracts associated with that family versus their frequency in abstracts treating unrelated proteins (see TF-IDF below, Document Classification). In essence, they created the first statistical BioNLP system able to annotate biological database records from scientific publications only.

Finally, the **natural language processing** (NLP) approaches developed earlier by linguists were introduced to biological problems by (Sekimizu et al. 1998). In their seminal work, they showed that purely linguistic techniques (specifically, *shallow parsing*) can be used to identify relationships between biological concepts in a sentence, using the verbs as relationship type. For example, the sentence "ETF binds to upstream sites in the p53 promoter." can be *parsed* to the relationship `'binds': ('ETF' → 'upstream sites in the p53 promoter')`. These four "founding" concepts have undergone significant developments and are now commonplace approaches in BioNLP, as we will see in the course of this work.

Document Classification

The first step in text mining commonly is the retrieval of relevant documents (IR). A classical IR example is FABLE, a system that retrieves lists of relevant articles for a human gene name (Fang et al. 2006), or XplorMed, that identifies articles similar to an existing set of texts (Perez-Iratxeta et al. 2003). Well-known examples of scientific IR systems are search engines provided by Google (Scholar) or PubMed, but also indexing engines built by the scientific community, such as HubMed (Eaton 2006) or EBIMed

(Rebholz-Schuhmann et al. 2007).

IR systems are often designed for keyword searches and at their core commonly employ an *inverted index* of terms, storing the total number of times a term appears in a document. That term count is subsequently normalized to a term frequency (TF). Documents containing a query term can then be retrieved in ranked order by dividing the term frequency[3] by the relative number of documents in which the term appears, the document frequency (DF[4]). Documents with high $TF \div -log\ DF$ (i.e., *term frequency times inverted document frequency*, or **TF-IDF**) scores for a term are the most specific documents for that (query) term (Jones 1972).

For example, the TF-IDF scores can be used to find the most similar documents for a query using the vector space model (Salton et al. 1975, Salton et al. 1983). Each word in a collection of documents represents a dimension, and each document is a vector of words. The TF-IDF values of the words in a document represent the coordinate values of the document vector in this "document space". Using, for example, cosine similarity as measure, these document vectors can be compared to a vector formed from the words in the query. The similarity scores identify the most relevant records for the query.

Apart from retrieving documents based on a query string, the vector space model can be used to cluster similar documents, too. For example, the above mentioned clustering of MEDLINE documents is based on these principles. A more modern approach for document classification is the use of Support Vector Machines (SVMs) with a **Bag-of-Words**[5] feature set and using the words' TF-IDF values a features (Joachims 1998). Due to the high dimensional input space, these Bag-of-Words are usually reduced with feature selection strategies such as *information gain* or the *chi-square statistic*. As shown in that seminal work, the sparsity of document vectors and the density of concept vectors (the Bag-of-Word, after feature selection) makes the SVM an ideal classifier for text categorization.

Named Entity Recognition

Information extraction (IE) analyzes documents, extracting factual (i.e., conceptual) knowledge from the text. Often, it is preceded by an IR technique to collect the relevant documents.

[3] For example, $TF := 1 + log\ n : t \in d$, where n is the number of times term t has been mentioned in document d

[4] I.e., $DF := |\{t \in d \forall\ d \in D\}| \div |D|$, defined as the number of times a term appears in any document divided by the total number of documents

[5] Essentially, the collection of all words in the documents.

The most obvious concepts found in biological publications are the genes and proteins (a type of biological *entity*) (Leser & Hakenberg 2005). Finding a specific entity type in a text - like gene names, diseases, chemical compounds, numerical quantities, etc. - is known as **named entity recognition** (NER) (Manning et al. 1999).

The most common learning algorithms used for NER are based on the principle of Markov chains, named after their discoverer, Andrey Markov (1906). Markov chains in essence are stochastic models of processes, formulated as a succession of states. The (random) process passes through these states, advancing from one state to the next, while these state transitions are associated with a probability. Two frequently used machine learning techniques for NER are based on the Markov principle, conditional random fields (CRFs) (Sutton & McCallum 2007) and hidden Markov models (HMM) (Rabiner & Juang 1986). The latter have been replaced by maximum entropy Markov models (MEMMs) to relax the strong independence assumptions made by HMMs (Ratnaparkhi 1996). The features used for these models are created from the morphological[6] and lexico-syntactic[7] properties of the words as well as the words themselves.

Well-known NER systems include ABNER (Settles 2005), the GENIA NER tagger (Tsuruoka et al. 2005)[8], BANNER (Leaman et al. 2008)[9], or OSCAR (Corbett & Copestake 2008)[10].

Entity Normalization

Recognized gene or protein names can be *normalized*, that is, assigned to their relevant database IDs. This process, known as **gene normalization** (GN), uses a lookup strategy of gene name-to-database ID relations stored in a *dictionary* (see Section 1.1.3) (Hirschman, Colosimo, Morgan & Yeh 2005). Commonly, dictionary entires are the (regularized) names or character-sequence patterns ("regular expressions") of genes mapped to their identifiers. The recognized gene names are *matched* to the dictionary, and thereby mapped to database IDs. It should be noted that GN is not necessarily preceded by a gene NER step described above - the lookups could be made directly on the text.

[6] use of upper- and lower-case, numerals, letters, symbols, etc.

[7] i.e., the lexical class/part of speech (noun, verb, adjective, etc.), and the grammatical number, tense, etc.

[8] Both the GENIA tagger and ABNER recognize mentions of DNA elements including genes, protein names, RNAs, cell lines, and cell types; both are CRF-based.

[9] BANNER recognizes gene and protein names only; it is CRF-based.

[10] Oscar3 recognizes chemical entities, i.e., chemical names, reaction names, and enzymes; it is MEMM-based.

The main handicaps in GN are resolving ambiguous genes with names assigned to multiple genes (homonyms), including the problem of recognizing the correct organism that encodes the gene (*species disambiguation*) (Fundel & Zimmer 2006). In addition, the name can be a homonym of a different concept than a gene, such as a referring to a gene family, a cell line, or even common words.

An example of inter-species ambiguity is the human gene PTEN; It has many synonyms[11], and is assigned to over 20 genes[12] from various species, not all of which are orthologs, either. An example of intra-species ambiguity is the acronym AAT1, assigned to three human genes (AAT1 itself, GPT, and C3orf15), or PAK1, assigned to two human genes (PAK1 and PKN1). A more involved example is the regular English noun "arm": (1) It officially is an abbreviation of the Drosophila armadillo gene, (2) but also is listed as a synonym for the gene product of the MED18 D. melanogaster gene (see SwissProt accession Q9XZT1), and (3) is listed as a synonym of the (hypothetical) protein of arm stored in TrEMBL (O46082). Together with the six regular English senses this word can take on when used as a noun[13], disambiguation is the complex matter of identifying the correct semantic role of a word given the context.

Knowing the relevant organism for the gene mapping is very important, too, considering that inter-species overlaps are generally rather common (Tuason et al. 2004, Chen et al. 2005). But making the correct organism mapping (i.e., identifying the relevant species for a gene/protein mention) is not straightforward, either. The relevant organism might not be quoted, which can make identifying the correct gene impossible, even for human experts. Otherwise, the organism might have to be inferred from a species-specific keyword (e.g., "murine"), a cell line mention (with all the issues as for GN itself), or the prefix of the gene name (h for human, m for mouse, At for A. thaliana, etc.).

Furthermore, the use of names changes over time and differs between literature and databases, too. Because names of genes are constantly evolving, it is hard to keep track of literature on a specific gene. The official gene nomenclature (see, for example, the HGNC (Seal et al. 2011) for human genes) might never get used in the literature, while not all names present in the literature have been recorded in repositories. In this vein, another possibility is that the full name is recorded in the databases, but the acronym is used in literature.

[11] BZS, DEC, GLM2, MHAM, MMAC1, TEP1, ...

[12] `http://www.ncbi.nlm.nih.gov/gene?term=PTEN%5BGene%20Name%5D`

[13] Arm can be used in the sense of referring to that body part, as a synonym for weapon ("to arms"), part of a seat ("armchair"), as a synonym for division, branch, or section ("an arm of Congress"), a part of clothing, or even used as an abstract concept similar to branch or limb, e.g. "an arm of the red sea".

Names from other biological areas overlap with gene names, such as names of cell lines ("CD4"), experimental methods ("SPR" for surface plasmon resonance), or diseases ("huntigton"), especially their respective abbreviations. Similarly, the name might also refer to a family name ("p53"), or a complex ("ARC").

As will be seen from the BioCreative results, these disambiguation issues are the reasons why GN remains a difficult problem, particularly on full-text and with no species restriction (i.e., if the system is not told a priori which target species it should use). For these reasons, "off-the-shelf" GN software is still scarce. GNAT (Hakenberg, Plake, Leaman, Schroeder & Gonzalez 2008) and GeneTUKit (Huang et al. 2011) are two recently published examples that are both developed by participants of the BioCreative efforts introduced in the next chapter, and Moara (Neves et al. 2010) is another.

Relationship Extraction and Hypothesis Generation

A primary goal of IE is the detection of biologically meaningful relationships between the entities found in the text. Well-known examples of systems extracting relationships are Chilibot (Chen & Sharp 2004) for graphing gene/protein interactions, or iHOP (Hoffmann & Valencia 2005) and GoGene (Plake et al. 2009) that can detect several more entity relations beyond genetic interactions.

To detect interactions between entities, a wide range of strategies are used, from simple co-occurrence of two entities within a sentence, to matching known patterns of expressions used to describe a relation between entities, and even analyzing the syntactic structure of a sentence to determine if two mentioned entities interact. Some of these techniques will be detailed in the course of this work, where they have been applied by BioCreative participants to identify protein interactions.

Going beyond existing relationships, KGT attempts to combine two or more relationships based on some correlation (usually, a "pivot" entity) to make novel discoveries. The classical reference for KGT is Swanson's work described already.

A very straightforward and accessible tool of this kind is Genes to Disease (G2D) (Perez-Iratxeta et al. 2005). This system aims at extracting candidate genes related to a genetic disease and a genomic region established by genetic linkage mapping. After determining a linkage mapping for a disease, an IE system identifies the gene(s) contributing to this phenotype via text mining the PubMed data. The user then is tasked with evaluating the potential novelty of the found relationships.

Another KGT example is Hanalyzer (Leach et al. 2009), a tool that was successfully

used to identify four novel genes involved in craniofacial development. Hanalyzer uses IE to build networks of ranked entity relationships, including inference techniques. A human expert then can evaluate these relations with respect to their novelty and relevance.

Ultimately, KGT systems are designed to infer knowledge by mining for non-obvious connections between the input data. It should be noted that KGT is mentioned for completeness, but will not be a focus of this work.

1.1.3 Resources

Scientific Literature

While PubMed is the single most universal text resource for BioNLP, these citation records suffer from an obvious shortcoming: Apart from title and abstract, no citation contains the full (body) text, images, or any other data content of the referenced literature. Furthermore, with the exception of a sample set of open access PubMedCentral articles[14] (about 340,000 articles) , even the open access publishers only allow manual (non-automated), single-article access (Carroll 2011).

Some publications - particularly those before the introduction of the World Wide Web - are distributed as scanned images only. PDF files and scanned text sources introduce noise already at pre-processing stages (i.e., the process of extracting the correct characters, paragraphs and headings in the right order, or filtering out unrelated text blocks, etc.) (Simske & Lin 2004). While we can only relate to our in-house experience with decompiling PDF files using open source tools such as Xpdf[15] or Apache PDFBox[16], these approaches produced results of varying quality. Using optical character recognition (OCR) systems seems to work better, e.g., (Stewart et al. 2007), and open source OCR tools available publicly are beginning to mature[17].

Due to this scarcity of full-text articles, providing text mining researchers with real-world article collections is an important contribution to BioNLP research. Freely available article collections can be exchanged between researchers and linked as publicly accessible data sets for their published work. For the evaluation of text mining approaches, these full-text articles collections are annotated with biological data. The BioCreative PPI corpora (e.g., (Leitner, Krallinger, Cesareni & Valencia 2010)) are examples of full-text collections annotated with PPI-relevant data.

[14] `http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`
[15] `http://foolabs.com/xpdf/`
[16] `http://pdfbox.apache.org/`
[17] e.g., the Tesseract OCR engine, `http://code.google.com/p/tesseract-ocr/`

Thesauri and Ontologies

External context knowledge for text mining is commonly provided through *controlled vocabularies*, *thesauri*, and *ontologies* (Spasic et al. 2005). They represent lists, trees, and relational graphs of concept terms, respectively, and map those concepts to a unique ID and description (Bard & Rhee 2004). In a broad sense, even databases - such as UniProt or Entrez Gene - represent controlled vocabularies or thesauri.

Mostly based on the need to agree on schemata for biological databases, thesauri and ontologies (such as the already described MeSH or the UMLS) were rapidly adopted by repository curators. Another well-known ontology is the Gene Ontology (GO) (Ashburner et al. 2000), commonplace for gene and protein annotation efforts such as the microarray analysis approaches reported above. But the GO has also been used in text mining research, e.g. (Raychaudhuri et al. 2002, Kim & Park 2004, Müller et al. 2004, Koike et al. 2005, Blake & Bult 2006, Zheng et al. 2006, Falcon & Gentleman 2007, Vanteru et al. 2008), to name a few. For example, the Abelson tyrosine-protein kinase 1 (ABL1) is annotated with the 45 GO terms[18]. These declare - for example - that ABL1 is associated with the "nucleus" (i.e., the associated *cellular component*), has "protein-tyrosine kinase activity" (i.e., the *molecular function* ABL1 is involved in), and plays a role in "regulation of transcription involved in S phase of mitotic cell cycle" (i.e., the *biological process*). In other words, ontologies provide standardized terms (*concepts*) that can be used to annotate biological entities with these concepts. However, less standardized annotations for genes exist as well, most prominently the GeneRIFs (Gene References Into Function) (Lu et al. 2007). GeneRIFs are Entrez Gene annotations submitted by researchers as short functional descriptions for genes they claim to be experts on.

Annotations of entities and their relationships occurring text are made with the identifiers of these ontology terms. Because of the use of standardized concept identifiers, these data become interoperable across databases, organizations or applications. For text mining, these resources are often used in the "reverse" direction. The names, terms, synonyms, and descriptions defined in the resource are used to detect the corresponding concept in the text. In the simplest case, this is approached by creating a *dictionary* of those names to the corresponding identifiers, much like the GN procedure described before.

Using these ontologies in BioNLP comes with its own set of issues. For example, GO was not designed for BioNLP tasks: It is unified for all organisms, while not all

[18] `http://amigo.geneontology.org/cgi-bin/amigo/gp-assoc.cgi?gp=UniProtKB:P00519`

processes and functions in the ontology exist in every organism; The limitation to two relations types can be insufficient for inferring relations (especially the broad use of "part of" relations) (Mungall 2004). As the GO is strictly separated into the three different hierarchies (cellular components, molecular functions, and biological processes), there are no relationships between terms in different categories (Ashburner et al. 2000). A general obstacle in using ontologies such as UMLS or GO and thesauri such as MeSH is the tentativeness of the mapping process between concepts and text spans, as the jargon used in scientific publications usually does not coincide with the descriptive expressions used in ontologies (McCray et al. 2002). In other words, identifying the correct ontology term from an expression that has no lexical match to the term is trivial for a (human) domain expert, while making these inferences automatically is not.

Corpora

Collections of sample texts (commonly, manually) annotated with database IDs, and/or concepts from ontologies and controlled vocabularies are provided as *corpora*. Some well known corpora are shown in Table 1.1. Notably, parts of the BioCreative corpora and the recent CRAFT corpus are collections of full-text articles.

These collections of articles provide annotations of concepts and entities (e.g., GO terms or genes) that are used to train and evaluate machine learning approaches for biomedical information extraction. While they are extremely labour-intensive to create, they are probably the most valuable resources in text mining. Essentially, for all IE methods presented above - NER, entity normalization, and interaction detection - one or more of the above corpora provides the training and evaluation data related to some biological topic (e.g., gene NER and normalization, PPIs, GO annotations, etc.). This lays the fundaments to allow researchers develop heterogeneous approaches, but directly compare these approaches using the same data sets.

Some of these corpora have been created almost a decade ago (GENIA and GENETAG), but remain in use to date, and some are continuously expanded (e.g., GENIA or BioCreative). Corpora with high annotation consistency are commonly called a "gold standard" in the NLP community, while collections with lower consistency, e.g., because they were created from automated annotation tools, are known as "silver standards". If a new method is developed, the performance of the method is evaluated on a corpus with corresponding annotations. To demonstrate that new approaches perform better than any previous know method, the corpora provide settings that make their evalua-

Name	Refence URL
AIMed	(Bunescu et al. 2005) ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/
BioCreative	(Hirschman, Yeh, Blaschke & Valencia 2005, Krallinger, Morgan, Smith, Leitner, Tanabe, Wilbur, Hirschman & Valencia 2008, Leitner, Krallinger, Cesareni & Valencia 2010, Arighi, Lu, Krallinger, Cohen, Wilbur, Valencia, Hirschman & Wu 2011) http://www.biocreative.org/resources/corpora/
BioInfer	(Pyysalo et al. 2007) http://mars.cs.utu.fi/BioInfer/
BioText	(Hearst et al. 2007) http://biotext.berkeley.edu/
CALBC	(Rebholz-Schuhmann et al. 2010) http://www.calbc.eu/
CRAFT	(Bada et al. 2010) http://bionlp-corpora.sourceforge.net/CRAFT/index.shtml
DrugDDI	(Segura-Bedmar et al. 2011*a*) http://labda.inf.uc3m.es/DDIExtraction2011/dataset.html
GeneTag	(Tanabe & Wilbur 2002, Tanabe et al. 2005) ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedTag/medtag.tar.gz
GENIA	(Kim et al. 2003) http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/
LLL	(Nédellec 2005) http://genome.jouy.inra.fr/texte/LLLchallenge/
Penn BioIE	(Mandel 2006) http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T21 http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T20

Table 1.1: A selection of BioNLP corpora.

tion transparent to the entire text mining community. And, because (most) corpora are freely available, other researchers can reproduce a published approach and independently measure the reported performance.

Databases

Sequence databases form another very important resource for text mining. For example, gene and protein names are commonly extracted from Entrez Gene (Maglott et al. 2011) and UniProt (Boutet et al. 2007) records. In addition, more specialized organism-specific databases such as FlyBase (Drosophila m., (McQuilton et al. 2011)) or TAIR (Arabidopsis t., (Lamesch et al. 2011)) can serve as sources of these names.

The more advanced gene normalization systems extract far more from the sequence databases than just names, however. To resolve the ambiguity of gene/protein homonyms, context information present in the text can be used. For example, GO term associations, sequence length, genomic location, etc., can serve as excellent "hints" to resolve these ambiguities.

Other database resources tapped for extracting context information are the NCBI Taxonomy[19] and NEWT (Phan et al. 2003) databases of species and common names, the OMIM (Hamosh et al. 2005) database of human phenotypes, the PubChem (Wang, Xiao, Suzek, Zhang, Wang & Bryant 2009) database of small molecules, or CLDB (Romano et al. 2009), a database of cell lines, among many others. These databases are used extract data that can help identify the respective entities (e.g., a specific chemical) in the text.

In addition to the sequence databases, the CLDB and NEWT or NCBI Taxonomy are commonly used resources for gene normalization (GN). These resources provide names and terms that can be used to detect the species mentioned in a text.

(Gene) Dictionaries and Normalization

As already described, dictionaries are collections of terms, expressions, or patterns that map to (ontology) concepts or entity identifiers, such as Gene Ontology term IDs, GeneIDs in Entrez, or protein accessions in UniProt.

Extracting names from databases or ontologies to create these dictionaries is a labor-intesive process. In particular, gene/protein name dictionaries require detailed, manual curation to be useful. While it is impossible to name all issues related to dictionary

[19] `http://www.ncbi.nlm.nih.gov/Taxonomy/`

assembly, we will name a few issues with creating gene/protein name dictionaries to highlight the complexity of this problem. For a more in-depth treatment of creating gene dictionaries, see (Koike & Takagi 2004) or (Tsuruoka et al. 2007). In addition, the BioCreative gene normalization tasks and their participants have provided a wealth of insight on the matter of protein/gene name dictionary construction, too (Hirschman, Colosimo, Morgan & Yeh 2005, Morgan et al. 2008, Lu et al. 2011).

The first step is usually **regularization** of the gene names, adding (possible) variations of names to the dictionary. Commonly applied regularization methods create spelling variations, such as adding/removing spaces, hyphens, or swapping letter cases. Another regularization technique refers to Greek letters and Roman numerals; These could be represented by their actual symbols, as Latin letters or Arabic numerals, or even spelled out. Long gene names might have been abbreviated by authors, therefore some researchers add acronyms of long names to their dictionaries[20].

All these variants and the original names then can refer to more than a single entity. Names that are common english words, single letters, or numbers need to be filtered or flagged as potentially ambiguous. Names that overlap with cell lines, experimental methods (particularly, their abbreviations), or diseases (e.g., "huntington") need to be treated with methods that can resolve the meaning of the particular mention when encountered in text. Sometimes, names overlap with gene family or protein domain names; These, too, need to be handled specifically; In BioNLP jargon, this process is is called **disambiguation**.

A common approach is to first use NER to detect gene names and then limit the dictionary lookup to the detected entities. The ambiguities can be resolved by many different approaches. The simplest disambiguation method is to resolve ambiguous names by defaulting to the more frequently encountered species for inter-species ambiguity or the main gene name for intra-species ambiguity. A more intricate approach is to detect context information, e.g., species mentions or genomic location, in close vicinity to the gene name that can help resolving these ambiguities. Given sufficient resources, some researchers even apply manual dictionary curation strategies to manipulate the entries in their dictionaries.

Furthermore, protein/gene names often have **affixes** with letters referring to species, experimental evidence, or other proteins. The prefixes "p-" (protein *or* phosphorylated), "h-" (human), "m-" (mouse/M. musculus *or* mammalian), "d-" (fruit fly/D.

[20] e.g., abbreviating "heat shock protein 90kDa alpha (cytosolic), class A member 1" (HSP90AA1) to HSP90α1

melanogaster), “c-” (celluar), and “v-” (viral), and the two-letter organism prefixes (e.g.,“At” for thale cress) are common, but inconsistently applied and in some cases ambiguous themselves (e.g., “m-”). Affixes of experimental origin (e.g., “3HA”, “anti”, “35S”, or “-/-”) can be found attached to gene and protein names, including the names of other proteins (His6, YFP, Myc, GST, flag, etc.) that are commonly used in experimental procedures (Krallinger, Tendulkar, Leitner, Chatr-aryamontri & Valencia 2010). These affixes can be used to resolve species ambiguity, or have to be masked if they are not to be normalized, as in the case of protein name affixes. If proteins applied in experimental procedures need to be extracted, some of the shown affixes could even be used as context information indicating their connection with experimental methodology.

Last, **matching** approaches need to be considered: One approach is to rigidly require full, exact matches. This removes some ambiguity at the cost of recall. Another is to check for partial matches of the names. For example, detecting shortened expressions, such as matching “Abelson murine leukemia 1” to its official name “Abelson murine leukemia viral oncogene homolog 1”. Finally, some researchers have considered fuzzy (approximate) string matching techniques. Fuzzy techniques are particularly useful for the more artificial names mentioned before and can detect similar variants of those names.

Suffice to say that the problems and solutions are too manifold to cover all of them. The interested reader is referred to the citations quoted at the beginning of this subsection. We hope to have given a glance the enormous depth of creating and using dictionaries, in particular gene dictionaries.

To conclude this section, tapping into biological repositories is an integral component of text mining approaches, but the proper use of these resources is often a research topic of its own. However, for more frequent concepts and entities, a few pre-built dictionaries exists. For example, the GNAT dictionary (Hakenberg et al. 2011) provides pre-compiled gene name mappings to Entrez Gene for ten different species. The BioThesaurus (Liu et al. 2006) provides the same mapping to UniProt. And the BioLexicon (Thompson et al. 2011) provides a complex compilation of concepts from many different resources, such as UniProt, ChEBI, NCBI Taxonomy, OMIM, InterPro, etc..

1.2 Community Challenges (in BioNLP)

1.2.1 Origins

We have shown that text mining has been used as the data (mining) source in many large-scale genomic and proteomic projects over the past years. Nonetheless, biologists, bioinformaticians, database curators, and even text mining experts all face difficult choices when intending to make use of existing approaches. Which methods are applicable and perform well for a particular task? Is there any existing software available for a method? What type of input data can be applied to the algorithm? What results quality can be expected? Which systems can use the same data or directly interoperate between each other? Etc. Regularly, these questions are resolved by a personal investigation of existing scientific research. Often, the performance of an approach will be the most relevant decision criteria. However, the reported evaluation results are not necessarily directly comparable between each other. This can be caused by incompatible evaluation metrics, the use of different corpora for the evaluation, or subtly different settings[21]. In an attempt to introduce independent evaluations as well as standardize evaluation strategies and metrics, scientific community challenges are held. Beyond the evaluation and standardization aspects, these challenges help foster development in their areas, because of their competitive nature and because they are drivers of community interaction and exchange.

Community challenges in computational biology have been called for several decades and even found their way to wet-lab molecular biology in 1997 (see Table 1.2). The Paracelsus challenge asked participants to change less than half of a protein's sequence to transform it into another globular protein. Before that, the first Critical Assessment of Techniques for Protein Structure Prediction (CASP) challenged computational biologists with the de-novo modeling of protein structure in 1994.

In BioNLP, community challenges were introduced first by a long-running data mining challenge, the **Knowledge Discovery and Data Mining (KDD) Cup**, in 2002 (Yeh et al. 2003). The first task, on IR, was to classify full-text papers (only available to challenge participants due to copyright issues) that are relevant for curators of FlyBase (the Drosophila genome DB). The second task consisted of annotating (with some classifier) a set of yeast genes with a binary label based on many different data resources

[21] E.g., comparing a cross-validation result to an evaluation that split the corpus in two, one for training and the other ("unseen") set for testing.

Name	Year∗	Refence†
CASP	1994	(Moult et al. 2011)
Paracelsus	1997	(Rose 1997)
CAMDA	2000	(Johnson & Lin 2001)
CAPRI	2001	(Wodak 2007)
GASP	2004	(Coghlan et al. 2008)
DREAM	2006	(Marbach et al. 2010)
CAGI	2010	not published to date (see `http://genomeinterpretation.org/`)

Table 1.2: A selection of community challenges in (computational; except Paracelsus) biology. Abbreviations: CAMDA (Critical Assessment of Microarray Data Analysis), CAPRI (Critical Assessment of Predicted Interactions), GASP (Genome Annotation Assessment Project), DREAM (Dialogue on Reverse Engineering Assessment and Methods), CAGI (Critical Assessment of Genome Interpretation). ∗: First year of the challenge. †: Latest reference.

(i.e., the features). One of these data resources was a set of about 15,000 MEDLINE abstracts, with the references to relevant abstracts assigned to each gene in the set. A notable added difficulty for this task was that participants were not informed about the meaning of the class labels. The labels later were revealed to identify genes that have an effect on the aryl hydrocarbon receptor signaling pathway in yeast when knocked out.

The **Text Retrieval Conference (TREC) Genomics** tracks, running from 2003 to 2007, started the next set of challenges, using PubMed abstracts (2003-2005) and full text (2006, 2007) as data resources (Hersh & Bhupatiraju 2003, Hersh & Voorhees 2009). The first track in 2003 asked participants to identify ranked lists of PubMed abstracts that contained biologically relevant information for a given gene or protein regarding their structural, genetic, and functional aspects. It should be pointed out that, while both these challenges' main text mining foci were document classification, the KDD Cup's second task had an IE aspect.

BioCreative (**Critical Assessment of Information Extraction in Biology**, starting in 2003) was the first call for the BioNLP community that challenged the participants with pure IE tasks (Blaschke et al. 2003). While the KDD Cup 2002 and the TREC challenges focused (mostly) on IR, the BioCreative tasks targeted (mostly) the issue of extracting information from text. The first BioCreative challenged participants with the functional annotation of genes from a collection of full-text articles, and will be described in detail below.

The next call for an IE task was the **Joint Workshop on Natural Language Processing in Biomedicine and its Applications** (JNLPBA, 2004), that focused on

NER (Kim et al. 2004), similar to the recurring gene mention NER tasks in BioCreative. Opposed to those BioCreative tasks, where the NER focused on gene names, participants at JNLPBA were asked to annotate several types of bio-entities (gene names, cell types, DNA elements, etc.) on approximately 3,000 MEDLINE abstracts.

In 2005, the **Learning Language in Logic** (LLL) challenge 2005 asked participants to extract gene interactions from about 130 sentences where only the genes were pre-labeled (Nédellec 2005). The correct interaction pairs had to be found by assigning linguistic information (lemmas, syntactic relations) to these sentences.

Another set of challenges, the **BioNLP Shared Tasks** (2009, 2011 so far) might be called the "descendants" of JNLPBA (Kim et al. 2009, Kim et al. 2011). While BioCreative focuses on the biological information extraction issues, these challenges embrace the linguistic aspects of IE: reporting (dependency) relations between detected entities (called "events" in these challenges). Similar to the recent BioCreative challenges, they mostly focus on PPI relationships, too (Saetre et al. 2010).

A very recent challenge in BioNLP was the **DDIExtraction** challenge in 2011[22], where the participants were asked to detect if drugs are described as interacting with each other. The detection of drug-drug interactions (DDI) is an important research area in patient safety, as these interactions can be dangerous health risks and increase health care costs[23].

It should be noted that there are many more challenges beyond the mentioned ones. There are, for example, biomedical NLP challenges on patient records (e.g., the Medical NLP Challenge[24]), or the **CALBC** challenges that are investigating the creation of a high-quality silver standard based on machine-annotations only (Rebholz-Schuhmann et al. 2011).

1.2.2 Structure of BioNLP Challenges

The basis of IE/IR-focused text mining challenges is a collection of texts with annotations, together forming a *corpus* (see Section 1.1.3 on Corpora). The common goal is to reproduce these annotations using only the text (and, if applicable, any external data sources) with an automatic approach. Due to the high standards applied in creating the annotations on a challenge corpus, the annotations are often termed the *gold standard*, and represent the ideal annotation result the automated systems should (re-) produce.

[22] see `http://labda.inf.uc3m.es/DDIExtraction2011/dataset.html`, publication pending
[23] `http://labda.inf.uc3m.es/DDIExtraction2011/background.html`
[24] `http://computationalmedicine.org/home-0`

A challenge usually consists of three major phases, the development/training, test, and evaluation phase. The challenge corpus is split in two, one set for training/developing the systems, and handed over to the participants several months before the actual challenge, together with the corresponding annotations. After this development phase, a short test phase (usually, a few days) represents the climax of a challenge: Participants receive the second half of the corpus, the test set, without the annotations (Annotations for the test set are not disclosed until after the test or evaluation phase.) Instead, during the test phase, participants are asked to produce annotations on the test set with their systems and submit these annotations to the organizers. During the evaluation phase, the organizers analyze the submitted results. Finally, the gold standard annotations on the test set are made public, and the systems as well as the conclusions from the evaluation are published.

1.2.3 The BioCreative Challenges

The aim of BioCreative tasks is to cater to the needs of biologists, bioinformaticians and database curators by fostering the development of information extraction approaches. The tasks should help gain a better understanding of different approaches and induce the development of new methods and applications. So far, the general topics of the past four challenges have focused on functional annotations of proteins (BioCreative I), protein interactions (BioCreative II & II.5), classification of protein interaction articles (BioCreative II, II.5, & III), the extraction of experimental evidence for protein interactions (BioCreative II & III), and creating human/graphical interfaces for GN tools (BioCreative III). See Table 1.3 for all main references.

Gene NER and GN in BioCreative

In addition, most challenges had two tasks focusing on core issue in BioNLP, namely the detection of protein and gene names (NER; gene mention task) and extraction of their identifiers (GN; gene normalization task).

An important initial step for many biomedical text mining tasks is the correct recognition of mentions of biological entities of interest, such as genes and proteins. This gene/protein named entity recognition was evaluated with the gene mention tasks of BioCreative I and II (Yeh et al. 2005, Smith et al. 2008). With F-scores[25] up to 90%, gene NER approaches have been shown to have become sufficiently robust.

[25] the harmonic mean between recall (i.e., coverage) and precision

Challenge	Paper Focus	Reference
BioCreative I	Overview	(Hirschman, Yeh, Blaschke & Valencia 2005)
BioCreative I	Gene Mention	(Yeh et al. 2005)
BioCreative I	Gene Normal.	(Hirschman, Colosimo, Morgan & Yeh 2005)
BioCreative I	GO Annotations	(Blaschke et al. 2005)
BioCreative II	Overview	(Krallinger, Morgan, Smith, Leitner, Tanabe, Wilbur, Hirschman & Valencia 2008)
BioCreative II	Gene Mention	(Smith et al. 2008)
BioCreative II	Gene Normal.	(Morgan et al. 2008)
BioCreative II	PPI	(Krallinger, Leitner, Rodriguez-Penagos & Valencia 2008)
BioCreative II	BCMS	(Leitner et al. 2008)
BioCreative II.5	Overview	(Leitner, Chatr-aryamontri, Mardis, Ceol, Krallinger, Licata, Hirschman, Cesareni & Valencia 2010)
BioCreative II.5	PPI	(Leitner, Mardis, Krallinger, Cesareni, Hirschman & Valencia 2010)
BioCreative II.5	Corpus	(Leitner, Krallinger, Cesareni & Valencia 2010)
BioCreative III	Workshop	(Arighi, Lu, Krallinger, Cohen, Wilbur, Valencia, Hirschman & Wu 2011)
BioCreative III	Gene Normal.	(Lu et al. 2011)
BioCreative III	PPI	(Krallinger et al. 2011)
BioCreative III	Interact.	(Arighi, Roberts, Agarwal, Bhattacharya, Cesareni, Chatr-Aryamontri, Clematide, Gaudet, Giglio, Harrow, Huala, Krallinger, Leser, Li, Liu, Lu, Maltais, Okazaki, Perfetto, Rinaldi, Saetre, Salgado, Srinivasan, Thomas, Toldo, Hirschman & Wu 2011)

Table 1.3: Main publications of the BioCreative challenges. Abbreviations: Gene Normal.: Gene Normalization (Task), Interact.: Interactive Demo (Task).

Biological sequence repositories (e.g., UniProt, or model organism databases) associate a unique identifier for each sequence on record. The focus of the gene normalization (GN) task was to return the list of gene identifiers mentioned in a text, in particular Entrez Gene IDs. In BioCreative I, the task was to return unique gene identifiers associated with gene products from a set of abstracts for specified model organisms (fly, mouse, and yeast) (Hirschman, Colosimo, Morgan & Yeh 2005). In BioCreative II, the task focused on human genes and their products (Morgan et al. 2008). During BioCreative III, the GN task had no limitation on the source organism and instead of abstracts, full-text papers were provided (Lu et al. 2011). All GN tasks used Entrez Gene as the target DB for mapping the gene identifiers. There were no gene mention or normalization tasks in BioCreative II.5.

The change of settings in the BioCreative III GN task made it significantly more difficult than its former invocations. The evaluation results showed a substantial decline in performance reflecting the increased difficulty of having to identify the relevant organism in BioCreative III *and* having to cope with full-text: Performance dropped from an F-score of 92% (best system result during the BC II GN task) to the comparable break-even point score[26] of 41% (best system result during the BC III GN task).

BioCreative I: GO Annotations

The Gene Ontology annotation task of BioCreative I addresses a common procedure of database curation workflows. In particular, the annotation of proteins with their cellular location, their molecular functions, and the biological process they are part of (see Figure 1.2).

Curators typically create functional annotations through associations of normalized entities to controlled vocabularies and ontology terms. The challenge task asked for the extraction of these associations, i.e., protein-GO term relations with the corresponding evidence passages. These relationships had to be extracted from full-text articles. Participants were asked to return triplets of protein ID (SwissProt accessions), GO term (IDs), and corresponding text passages from a set of about 1000 full-text articles (80% used for training, 20% for testing). These tasks cover all but the (green) article detection step in Figure 1.2.

Professional GOA (the GO Annotations DB, (Barrell et al. 2009)) curators created the annotations, and the DB did not publish their annotations for the 200 test set

[26] measured at the position in a ranked list of results where precision and recall are (almost) equal

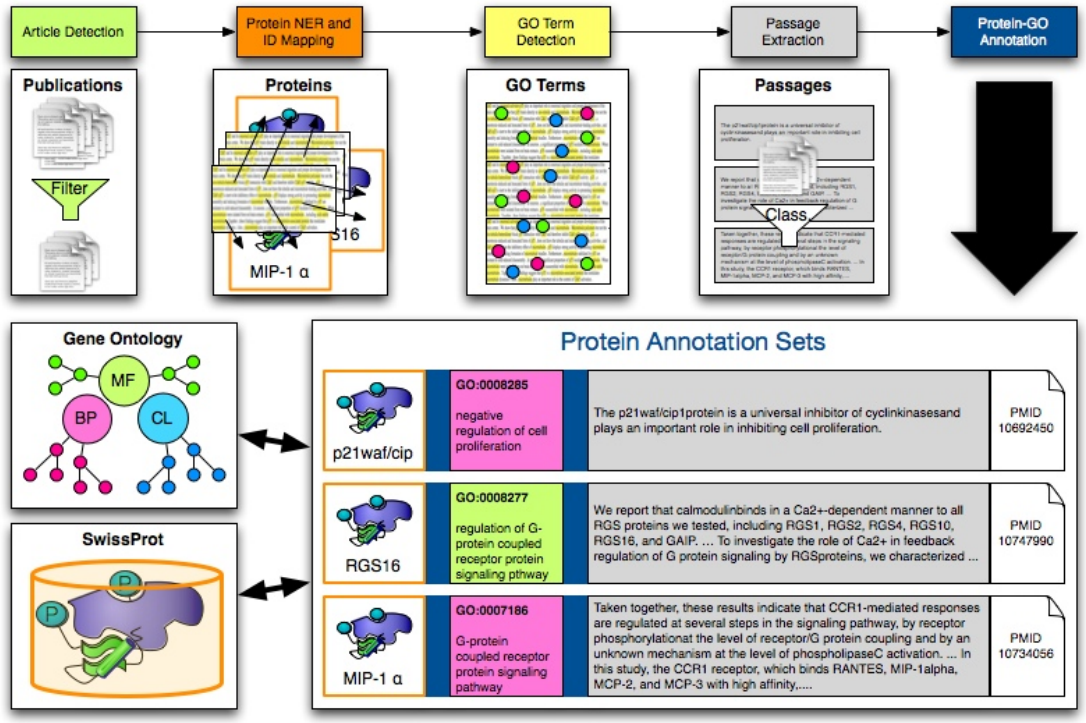

Figure 1.2: BioCreative I: Extraction of functional annoations (as GO terms) for proteins from the biomedical literature. The first step in a BioNLP pipeline would be the identification of articles containing functional descriptions of proteins (green; not part of this challenge). The proteins mentioned in the article had to be detected and mapped (i.e., normalized) to the given protein SwissProt identifiers (orange). Then, expressions in the abstracts for Gene Onotlogy concepts had to be detected and mapped. The relevant passages containing normalized proteins and GO terms are extracted by a classifier (grey). Finally, the right triples of protein ID, GO term, and passage for each article are reported by the system (dark blue). This results in annotations of text passages containing functional descriptions of proteins linked to their DB IDs and the corresponding GO terms (bottom).

articles until after the challenge. The participants received the training articles and had approximately three months time to develop their systems. Then, they had five days between receiving the test data and returning their results during the test phase.

The task was split into two: For the first sub-task (the *evidence extraction* task), participants were given a list of SwissProt accessions and the list of GO term IDs to annotate on each protein in the article. In other words, the challenge was to extract the relevant evidence passages and choose the right accession, term pairs for the passage. For the second, more difficult sub-task, participants were only provided with the number of GO terms that should be returned for each protein (i.e., an *annotation extraction* task).

The results in both cases had to be protein, GO term, passage triples. The pas-

sages had to coincide with the passages selected by professional GOA curators for the annotation. However, even the easier first (evidence extraction) task is non-trivial: Systems have to recognize the protein mentions and attempt to normalize those mentions to identify the correct protein. They have to identify expressions in the text that can be mapped to a GO term. This involves most of the challenges discussed in Gene Normalization (Section 1.1.2) and Dictionaries (Section 1.1.3). The passages can be a single or multiple sentences. To detect multiple sentences, a system has to be able to infer that consecutive sentences refer to the same protein or concept. However, natural language allows referring to an object in a previous sentence by using only a determiner or pronoun. Therefore, to detect all relevant sentences, a system ideally would need to be able to trace references (determiners, pronouns) to an object discussed in an earlier sentence, identifying the referent (a process called *co-reference* or *anaphora resolution*). However, this co-reference resolution is one of the more challenging issues in NLP in general (Dai, Chang, Tzong-Han Tsai & Hsu 2010).

For the evidence extraction task, the best precision was achieved by Chiang et al (80%, although at less than 4% recall), and the best recall system (Krallinger et al.) reached 28%, at a similar level of precision (29%) and achieving the best harmonic mean (28%) from both precision and recall (i.e., F_1-measure or F-score, see Section 3.2.3) of all participants. The results for the full annotation extraction task (blinded GO term IDs) were considerably lower, with the best precision of around 32% (at 3% recall, Chiang et al.) and the best recall not surpassing 7%.

The BioCreative I results conclusively demonstrated that concepts for cellular components and molecular function are easier to trace in text than biological processes. This can be explained by the variety of possible descriptions for biological processes in literature versus the more concise terms used for function and component/location concepts. In a number of cases, the official GO term for a biological process was not present in the article, while function and component terms had less divergence in this respect. A discriminative factor between teams was the protein normalization performance; It was shown to impact the overall result.

The inter-annotator agreement of curators from the GOA database, if using the same exact GO term and passage matching of the annotations, showed that the system results are not that far behind, as DB agreement scores were below 40% Joint Probability of Agreement (JPA[27], a measure similar to an F-score). In non-technical terms this means

[27] see Section 3.2.4 for inter-anntoator agreement measures

that even curators have problems agreeing on the same text passages that should be judged relevant for an annotation. The effect of BioCreative I is visible in its aftermath; It fostered the development of various tools on this biologically relevant task, such as the already mentioned GoGene system or the GoAnnotator (Couto et al. 2006).

1.3 The BioCreative PPI Challenges

1.3.1 Overview of the BioCreative PPI Tasks

Introduction

For the BioCreative II, II.5, and III challenges, the main focus was on protein-protein interactions (PPIs). Several databases are engaged in manual annotation of PPIs from the literature, including MINT (Licata et al. 2011), IntAct (Hermjakob et al. 2004) and BioGRID (Stark et al. 2011), jointly known as the IMEx Consortium[28]. This ensured that all PPI data used for the BioCreative challenges meet the standards of the PPI bio-curation community (Kerrien et al. 2007). Text mining of PPI-relevant data from the scientific literature has been a major focus of recent research (Jose et al. 2007).

The PPI challenges evaluated the performance of automated systems on tasks mimicking the manual literature curation workflow (Chatr-aryamontri et al. 2008, Chatr-Aryamontri et al. 2011). The four BioCreative PPI tasks we will focus on in this work are shown in Table 1.4. An additional task during BioCreative II was the retrieval of evidence sentences for the PPIs as defined by the DB curators (Krallinger, Leitner, Rodriguez-Penagos & Valencia 2008).

An important fact to note is that for BioCreative II, there was no separate protein normalization task. Just as BC I and III, BC II already had a gene normalization task. Therefore, for this work, all interaction pairs reported by the BC II systems were decomposed to single, unique interactions to evaluate the normalization performance. On the other hand, for BioCreative II.5, protein normalization was its own, separate task. Although systems for these two tasks should - at least in theory - report the same proteins, this significant difference in settings needs to be acknowledged.

Not all mentioned PPIs in an article are relevant for database curation; Only interactions for which the authors provide experimental evidence are curated by PPI databases. This issue is addressed by two tasks:

[28] http://www.imexconsortium.org/

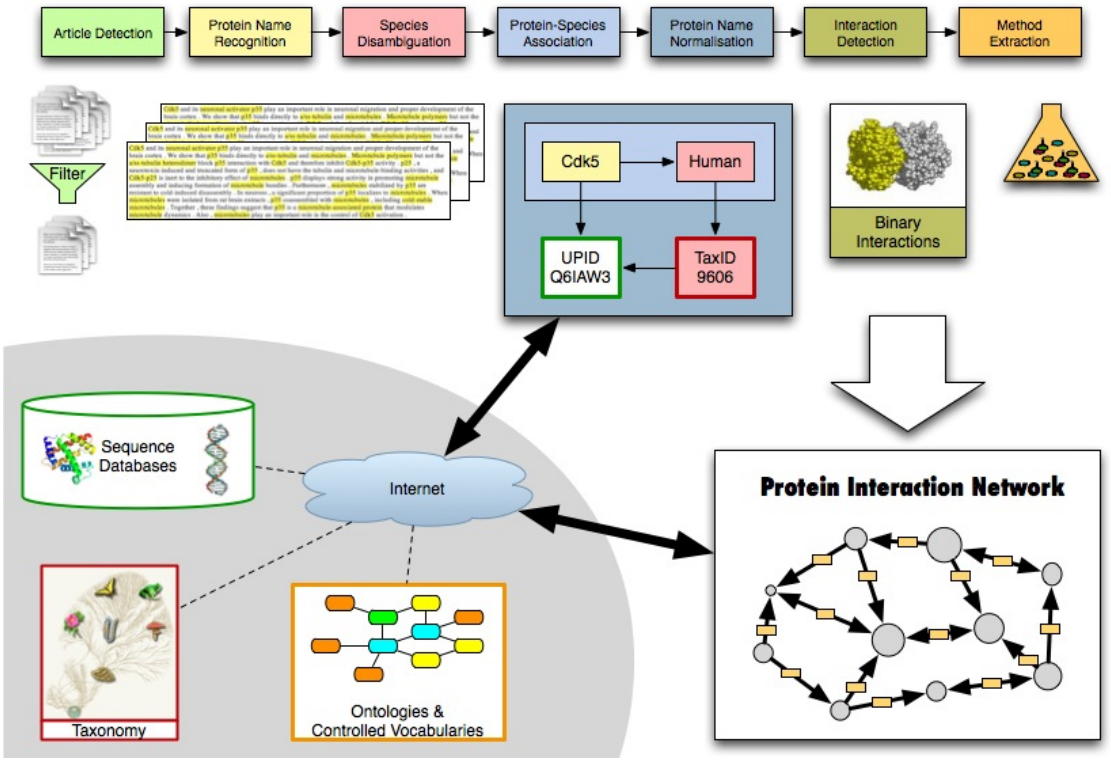

Figure 1.3: BioCreative II-III: Extraction of PPIs and corresponding experimental methods from the biomedical literature. The first step in curating PPI information is the retrieval of PPI-information containing articles, modeled as the article classification task (green). For these PPI articles, the proteins used in an experimental, PPI-relevant context need to be identified and their UniProt accessions have to be mapped (yellow, red and blue steps); This includes recognizing the relevant organism for the proteins ("species disambiguation"). These four steps together represent the protein normalization task. Next, the correct interaction pairs for the normalized proteins need to be detected (interaction detection task), and finally, the experimental methods extracted (method extraction task). Note that method extraction is shown as the last step, however, it could be placed even before the protein normalization, ensuring the extracted proteins are part of an experimental setting.

Name	Challenges	Description
Article classification	II, II.5, III	Classification and ranking of abstracts (BC II and III) or full-text articles (BC II.5) according to their likelihood of containing experimentally verified PPI data.
Protein normalization	II, II.5	Annotating and ranking the protein identifiers that have experimentally backed PPI evidence.
Interaction detection	II, II.5	Reporting ranked lists of protein interaction pairs that have been demonstrated by experimental evidence.
Method extraction	II, III	Extracting the experimental methods used to detect the PPIs in the article as PSI-MI ontology terms.

Table 1.4: BioCreative PPI tasks. Names and descriptions of the four tasks considered in the context of this work, and the corresponding BioCreative challenges where the task was present. Note that all but the first task were based on full-text articles. Proteins are reported as UniProt accession IDs.

1. The article classification task, asking participants to identify articles describing PPIs verified in an experimental setting.

2. The experimental method extraction task addresses the issue by asking participants to explicitly report the experimental procedure used, as PSI-MI term IDs.

The PSI-MI is an ontology maintained by the Proteomics Standards Initiative (PSI) and aims to facilitate data comparison, exchange, and verification for molecular interactions (MI) (Kerrien et al. 2007). It is the default ontology used for annotating PPIs by the IMEx PPI databases (Orchard et al. 2007). One branch of the ontology contains a thesaurus of experimental methods that are used to verify protein interactions. The ontology term IDs found in this branch were the annotation targets for the method extraction task. The protein normalization and interaction detection tasks are specifically limited to only those interacting proteins in the article that had been detected using some experimental PPI detection method. In other words, a system extracting a protein or PPI that is mentioned in the article, but had not been experimentally verified in the context of the publication, is treated as committing to a false positive error.

BioCreative II.5 and III were scientific community challenges held online[29]. I.e.,

[29]Participant were allowed to submit offline results, too, if they could not provide an online service. However, they were encouraged to use the online scenario and provided with the technical assistance

participants were asked to provide their systems as web services. During BioCreative II.5, the systems were not only tested online, but participants completed the training/development phase online, too. The environment for running this online evaluation was provided through the BioCreative Meta-Server (BCMS) (Leitner et al. 2008). This platform was created after the BioCreative II challenge in collaboration with its participants, and represents the first meta-server platform for BioNLP. The platform was then modified to allow participants of BioCreative II.5 and III to command and control the interaction of the server with their text mining systems. The BCMS also provides rudimentary insights on issues regarding interoperability in text mining: It represents an attempt to unify in- and output of text mining systems and allowed us to measure online annotation times. Because of the use during the challenge, it was possible to monitor the behavior of an distributed BioNLP service in a realistic setting.

1.3.2 An Interactive Demo Task

One of the main parts of BioCreative III was an interactive systems demonstration task (Arighi, Roberts, Agarwal, Bhattacharya, Cesareni, Chatr-Aryamontri, Clematide, Gaudet, Giglio, Harrow, Huala, Krallinger, Leser, Li, Liu, Lu, Maltais, Okazaki, Perfetto, Rinaldi, Saetre, Salgado, Srinivasan, Thomas, Toldo, Hirschman & Wu 2011). Its focus was the development of curation/annotation interfaces for the BioCreative III gene normalization task. The goal was to design user interfaces that can assist the curators in their workflow. Instead of scoring participant results using an evaluation metric, this task was designed as an exploratory, qualitative experiment to be further developed in future BioCreative challenges. A user advisory group consisting mostly of professional database curators participated during the design and assessment phase, addressing the utility of interfaces for the extracted GN information. Given that recognizing and normalizing gene mentions requires a major proportion of a curator's time (Chatr-aryamontri et al. 2008, Leitner, Chatr-aryamontri, Mardis, Ceol, Krallinger, Licata, Hirschman, Cesareni & Valencia 2010), designing interfaces for the recurring GN task of BioCreative is an objective with obvious benefits for the bio-curation community. Another target community would be journals, or, more precisely, their authors (Leitner, Chatr-aryamontri, Mardis, Ceol, Krallinger, Licata, Hirschman, Cesareni & Valencia 2010). In this scenario, experimental user interfaces for text mining systems could aid scientists in the annotation of their publications with relevant database iden-

required to do so.

tifiers. Put differently, the insights gained from this task provided the first pointers as how to harmonize text mining interfaces so that they optimally serve the purpose of creating author- or curator-based annotations.

In total, six systems addressing the interactive task were presented. All processed and displayed gene normalization results for the same set of articles. Three articles were chosen for the final assessment by the user advisory group members, each member reviewing one of the six systems. One common issue reported by the reviewers was the suboptimal system performance of the text mining systems on the GN task itself[30]. Another important issue brought up by the advisory group members was that interactive systems should display contextual information about a normalization; For example, synonyms on record for the gene, known genomic position, functional information, or sub-cellular location - in essence, any information that could assist a human in the task of identifying the correct gene record from the list of results. Often, one or a few of the sentences/passages mentioning the same gene within an article might contain highly specific information that can be used to identify the correct normalization. Curators reported that it would be useful if a system could group all sentences or passages where one gene is mentioned, and interactively update that passage collection with new (or corrections of automated) gene ID assignments made by the curator in the interface.

1.3.3 The BioCreative Meta-Server

During the BioCreative II workshop, the possibility of motivating participants to provide their methods online and make the data generated during the challenges publicly available lead to an initial implementation of the first text mining meta-server, the BioCreative Meta-Server (BCMS) platform (Leitner et al. 2008). The BCMS is a prototype system able to collate multiple predictions from various annotation servers based on web services and accessible to end users via a web interface[31].

This server then was used during the BioCreative II.5 and III challenges to evaluate the performance of the systems on the PPI tasks in an online environment. This made it possible to monitor the feasibility of providing online annotation services, observe if there was any negative impact from requiring online analytical processing ("OLAP") by comparing the online to the offline results, and investigate possible issues on interoperability in BioNLP.

[30] see Section 1.2.3, "Gene NER and GN in BioCreative" above
[31] `http://bcms.bioinfo.cnio.es/`

1.3.4 The PPI Tasks Setting

PPIs are involved in virtually every cellular process and errors in this interaction network often are the root cause of diseases (e.g., (Kiel et al. 2011, Oti et al. 2006)). Physical PPIs are the foundation of the interactome (the mesh of all molecular interactions in cells). Large-scale approaches such as yeast two-hybrid screening have been used to map large parts of the PPI network, and sequence, structural, and evolutionary information has been used in computational biology to predict these networks (see Section 1.1.1).

Biology constitutes an empirical, evidence-based science, and biologist evaluate the quality of biological data using the experimental methods applied to generate them. This means that the experimental procedure used to verify a PPI is a central element of interaction annotations. Each experimental method has different qualities that lead to certain approaches being more robust than others (i.e., produce less false results), usually at the expense of the possible number of interactions that a single experiment can cover (Lalonde et al. 2008) (among others, see Section 1.1.1). This experimental evidence is used by biologists to infer a sense of "trust" in a given PPI annotation.

These PPIs are catalogued in a number of different databases, such as HPRD, MINT, IntAct, or BioGRID. To compare annotations from different databases, both a standard annotation format - the MIMIx standard (Orchard et al. 2007) - and a controlled vocabulary - the PSI-MI ontology (Kerrien et al. 2007) - for the categorical annotations, such as the experimental methods, have been developed to coordinate these various databases. The manual extraction of PPIs from the scientific literature by professional curators follows strict guidelines to ensure uniform results and quality. The PPI databases are used by researchers to work on issues ranging from understanding the detailed interaction network of a small set of proteins to the reconstruction of the entire network for an organism; These interactomes are the key prerequisite for modeling cell physiology, and are a necessary milestone for moving systems biology into mainstream health care (Weston & Hood 2004).

Note that published PPI data not always arises from interactions between proteins of the same species. Cross-species interactions can be found in host-pathogen scenarios, such as viral and bacterial proteins interacting with a mammal protein, or they might originate from cloning and tansfection experiments using a cross-species target, including experiments to define the common structure and activity of orthologs. The corpora of the BioCreative challenges therefore contain a minor, but relevant[32] proportion of cross-

[32] on average, about 10-15% of all interactions

species interactions.

At current funding levels, PPI databases can only keep up with about 20% of the yearly interaction data that is published by the biosciences: The IMEx consortium, which represents the majority of all PPI data, curated *roughly* two thousand[33] of the estimated 10,000 PPI articles in 2009[34]. In other words, the larger part of the generated scientific data remains "dormant" in written text, because it is neither trivial to retrieve and extract, nor is this format accessible for large-scale data manipulation as would be required by systems biology or other data integration approaches. Furthermore, at least preliminary results measuring curation times using (text mining-) assisted curation have shown that these environments can decrease curation times (Alex, Grover, Haddow, Kabadjov, Klein, Matthews, Roebuck, Tobin & Wang 2008).

1.3.5 The FEBS SDA Experiment

In addition to the possibility of improving curator workflow, the IE systems for PPIs could improve the quality and/or time of generating Structured Digital Abstracts (SDAs). The SDA experiment is an initiative launched by the FEBS (Ceol et al. 2008) journals (Federation of European Biochemical Societies). SDAs are digital abstracts that can be interpreted by both human and machines and are provided by the FEBS Journal and FEBS Letters research articles attached to the "traditional" abstract (Ceol et al. 2008).

In particular, SDAs provide annotations describing PPIs that are reported in the publication, together with the experimental evidence and the interaction type for each interaction. The BioCreative II.5 corpora is largely based on the data generated by this SDA effort. In principal, SDAs are added to the publications by database curators of the MINT PPI database. However, for the period of nearly one year (during 2008), FEBS Letters asked authors of PPI-reporting papers to provide their own annotations for the interactions along the lines of the MIMIx recommendations. The goal of this experiment was to review of the impact of these author annotations on curation throughput and quality.

Therefore, trained curators from the MINT PPI database then examined these annotations and based their own annotations on this author data. In addition, later these authors were asked to respond to a questionnaire to measure the perception of this effort

[33]IMEx curation estimates for 2009 were made by personal communication with IMEx members (Gianni Cesareni and Andrew Chatry-aryamontri).

[34]The approach used for estimating the published PPI articles in 2009 is described in the Appendix, Section A.2.

and their willingness to cooperate. This SDA experiment resulted in an additional data set that was evaluated during BioCreative II.5: in addition to challenge participants and the IMEx curator data, these author annotations were used to directly compare all three annotation sources (i.e., authors, curators, and IE systems). This joint BioCreative/FEBS Letters SDA experiment evaluated the possibility of including authors in the annotation process, compared their results to automated systems, and questioned them about aspects of annotating their own publications (Leitner, Chatr-aryamontri, Mardis, Ceol, Krallinger, Licata, Hirschman, Cesareni & Valencia 2010).

1.4 Author Contributions and Responsiblities

The last three BioCreative challenges were collaborations between researchers at the European Bioinformatics Institute (EBI), the MITRE Corporation, the National Center for Biotechnology Information (NCBI/NIH), the Spanish National Cancer Research Centre (CNIO), and the Universities of Colorado, Delaware, Jena, Standford, and Tor Vergata (Rome). In addition, professional bio-curators of the MINT, IntAct, and BioGRID IMEx PPI databases created the annotations for the gold standards.

The author of this work was responsible for the following parts of the BioCreative challenges:

- During BioCreative II, he acted as co-organizer, where he had a minor role in the evaluation of the PPI tasks and organized the publications of all tasks.
- During BioCreative II.5, he acted as the main organizer, and had a major role in all aspects of the challenge.
- During BioCreative III, he acted as co-organizer, where he had a minor role in organizing the challenge, workshop, and publications, and took part in the evaluations of the PPI tasks.
- During BioCreative II.5 and III, he was responsible for coordinating and supporting online teams during the implementation of their web services.

He declares authorship of the following parts of the work:

- The thesis author is the developer and maintainer of the BioCreative evaluation software and website.

- The thesis author is the designer and developer of the BioCreative Meta-Server.
- The thesis author devised the FAP-score, and to his best knowledge no prior work exists that combines F-measure and Average Precision.
- The thesis author created the ensemble approach presented in the discussion of this work.
- The thesis author is responsible for all evaluation results and the assessment of participant systems presented in this work.

CHAPTER 2

Objectives

The main objective of this work is the assessment of current text mining approaches for protein-protein interaction (PPI) extraction. The study should provide answers relating to the ability of text mining to support database professionals in their curation work and/or point out feasible avenues to integrate authors in this process.

We have organized three BioCreative community challenges (BC II, II.5, and III) on PPI extraction to investigate these scenarios. We will present our approaches of assembling the corpora (Section 3.1), present an evaluation of the corpora consistency (Section 4.1), and discuss the real-world properties of these corpora (Section 5.1).

To measure the performance of participants, we have compared traditional evaluation metrics (Section 3.2). We will present a new, combined metric that can evaluate ranked results (FAP-score, Section 3.2.5), but penalizes results with many irrelevant annotations without the need of an artificial threshold (Section 5.2).

The thesis presents our results of evaluating the participant submissions on the following four tasks:

1. Classification of PPI-relevant articles (article classification task, Section 4.3.1).
2. Mapping of mentions of interacting proteins to database identifiers (protein normalization task, Section 4.3.2)
3. Detection of protein interaction pairs (interaction detection task, Section 4.3.3).
4. Extraction of experimental methods used to detect these interactions (method extraction task, Section 4.3.4).

The evaluation settings and materials are presented in Section 3.3. In the Discussion, we will outline the issues and capabilities of the 52 participant systems (Section 5.3).

In a real-world scenario, automated PPI annotations would be provided online. A simulation of an online scenario was designed for the challenges. We implemented a meta-server that collates the annotations of participants using web services. In this work, we will outline the design of this meta-server (the BioCreative Meta-Server, BCMS, Section 3.5), present the results of applying it during the community challenges (Sections 4.2 and 4.4), and discuss its contributions and utility (Section 5.4).

Finally, we will connect all results to discuss possibilities of creating a more effective approach for author- or curator-based PPI annotations by incorporating text mining systems into their workflows (Section 5.5).

CHAPTER 3

Materials and Methods

3.1 BioCreative PPI Tasks Corpora

3.1.1 BioCreative II Evaluation Corpora

The 1108 positive full-text articles (with annotations) used for the BioCreative II protein normalization, interaction detection, and method extraction task evaluations were all manually curated by MINT and IntAct curators. The full-text articles were supplied to the participants in four versions: 1) as HTML and 2) PDF, and as their corresponding plain text versions converted by the Unix tools 3) `html2text` and 4) `pdftotext`. The 3,874 positive and 2,298 negative abstracts for the article classification task consisted of titles inspected by the DB professionals during their curation process.

In addition to these training and test set abstracts, participants of BioCreative II received a set of "likely positive" abstracts for the article classification task. This "likely positive" set consisted of 18,930 abstracts curated by other PPI databases (e.g., BIND, HPRD, or BioGRID). As these databases might have different curation guidelines than MINT and IntAct, these papers are not part of the official positive collection of training set articles.

The article classification task gold standard is a table of PMIDs annotated with a Boolean value (true, false) denoting the abstract's relevance for PPI curation. For the protein normalization, interaction detection and method extraction tasks, each article is annotated by PMID with one or two UniProt accessions (single accessions for protein normalization and pairs for interaction detection) or a PSI-MI IDs (method extraction) per annotation. In total, 25,102 curated abstracts (including the "likely positive" training set) were provided for the article classification task, 974 papers were used for the normalization and interaction tasks, and all 1108 papers for the method extraction task. Papers that did not mention all relevant proteins in the text were removed from the normalization and interaction article test sets.

Furthermore, for 338 of the test set full-text articles, the MINT and IntAct curators provided the evidence sentences that they had judged as the most informative sentences for the annotated interactions ("interaction sentences"). These evidence sentences were used for the interaction sentence classification task (not described in this work; refer to (Krallinger, Leitner, Rodriguez-Penagos & Valencia 2008)). They could be used to train future classifiers, because they indicate the most relevant passages for feature extraction.

3.1.2 BioCreative II.5 Evaluation Corpora

The entire BioCreative II.5 article set consists of 1190 full-text XML files, all from FEBS Letters, and were all curated by MINT curators. 124 of these files were classified as positive articles containing PPI descriptions, of which 122 have interaction annotations. The remaining 1066 papers are negative articles containing no PPI information. The 122 articles that have PPI annotations[1] were used for the normalization and interaction tasks.

Due to the fact that the author annotations had already been made public at the time of the BioCreative II.5 challenge, the articles annotated by authors could not be used as test set. Instead, these author-annotated articles were part of the training set and a set of articles from the year before (2007, without public SDAs) was used as test set. In addition, nine annotated article in the test set were from the years 2002-2006, and had no PPI annotations in any database.

Several different sets of articles were created for the evaluations:

Authors corpus: The corpus to evaluate author performance consisted of the 51 articles from 2008 for which participating authors (57 in total) returned evaluation-relevant PPI data. 56 authors returned annotations, while one author did not, and 5 of the 56 authors did not return annotations for protein identifiers or pairs.

Curator corpus: The curator corpus consisted of 42 articles for which the curators generated annotations not basing their work on the author data, 38 of which overlap with the *author corpus* (i.e., without the four articles not annotated by authors), and 33 of which in turn form the *overlap corpus*. Furthermore, 41 of these 42 articles were used for calculating the IAA (inter- and intra-DB agreement, with 21 and 20 articles respectively).

Author-based curator annotation corpus: For the evaluation of curator performance basing their work on author annotations, 55 articles were used. These articles cover the 51 author-annotated articles plus four of the missing articles in the author corpus for which authors had submitted annotations without PPI pair or identifier information. These four articles nonetheless had been used by the curator for generating the author-based annotations.

Overlap corpus: The overlap corpus consists of 33 articles that were annotated in common by a curator, the authors, and the systems (I.e., 42 articles from the curator

[1] Two articles in the test set were initially classified (wrongly) as negative by curators and re-classified later by the BioCreative organizers during the evaluation phase.

corpus minus four without author annotations and minus five articles which do not overlap with the BioCreative II.5 training set.) This corpus is used in the overlap evaluation.

In addition to these "special" evaluation corpora, the training and test set are formed as follows:

Training set: The training set consisted of 61 FEBS Letters articles from 2008, mostly containing the publicly available SDAs created by authors and/or curators. The negative articles were a random selection of 534 other FEBS Letter articles from 2008.

Test set: The test set consisted of 61 (normalization and interaction tasks) and 63 (article classification task) FEBS Letters articles, and 52/54 of these were from 2007, but also included nine articles from the period of 2002-2006 that had no annotations in any PPI repository at the time of the challenge. This corpus has no overlaps with any of the other corpora. The negative articles for the article classification task were a random selection of 532 additional FEBS Letters papers from 2007.

3.1.3 BioCreative III Evaluation Corpora

For the BioCreative III *article classification corpus*, in part, this work was distributed to biologists especially trained for this task ("expert curators"), to reduce the workload on the databases. The corpus was manually labeled by database curators and trained domain experts using the MyMiner[2] "File Labeling" feature (Arighi, Roberts, Agarwal, Bhattacharya, Cesareni, Chatr-Aryamontri, Clematide, Gaudet, Giglio, Harrow, Huala, Krallinger, Leser, Li, Liu, Lu, Maltais, Okazaki, Perfetto, Rinaldi, Saetre, Salgado, Srinivasan, Thomas, Toldo, Hirschman & Wu 2011). MyMiner records both the label and the labeling time. The training set consists of 6,280 abstracts; A balanced set of 2,280 relevant and non-relevant PPI abstracts, and a development set of 4,000 where 17% of the abstracts are PPI relevant. 15% of the abstracts in the test set are relevant (out of 6,000).

The *method extraction corpus* training set spans 2,590 PDF articles from 87 different journals that are also available as plain text, extracted using the Unix `pdftotext` tool. The articles all had been curated by MINT and BioGRID curators. For the test set, participants received 305 articles, but only 223 of these were annotation-relevant (the other 82 articles were not used during evaluation and had only been added during the test phase to ensure participants were not manually generating their annotations).

[2] `http://myminer.armi.monash.edu.au/`

Condition	True Instance	False Instance	**Measure** ↓
Positive Result	*True Positive*	*False Positive*	Precision
Negative Result	*False Negative*	*True Negative*	Negative Precision
Measure →	Recall (Sensitivity)	Specificity & Fall-out	Accuracy

Table 3.1: Relationship between instance state (true/false instance), classification results (positive/negative result), and evaluation metric (measure).

3.2 Evaluation Metrics

3.2.1 Nomenclature

FN false negatives - incorrect negative classification results (type II errors)
FP false positives - incorrect positive classification results (type I errors)
TN true negatives - correct negative classification results (correct rejection)
TP true positives - correct positive classification results (correct ***hit***)
a accuracy
f fall-out (false positive rate, false alarm rate)
n negative precision (negative predictive value)
p precision (positive predictive value)
r recall (coverage, sensitivity, true positive rate, hit rate)
s specificity (true negative rate)

Classifier evaluation metrics are commonly based on a small set of basic statistical measures of the performance of a system. The relationship between (gold standard) annotations on the test instances, the classifier's results, and these measures can be categorized as shown in Table 3.1.

3.2.2 Basic Measures

The basic metrics are recall, precision, negative precision, specificity, fall-out, and accuracy.

Recall r (also known as coverage, sensitivity, true positive rate, or hit rate) is the percentage of correctly labeled positive results over all positive cases.

$$r := \frac{TP}{TP + FN} \tag{3.1}$$

It is a measure of a classifier's ability to identify positive cases.

Precision p (a.k.a. positive predictive value) is the percentage of correctly labeled positive results over all positive labeled results.

$$p := \frac{TP}{TP + FP} \tag{3.2}$$

It is a measure of a classifier's reproducibility of the positive results.

Negative precision n (a.k.a. negative predictive value) is the percentage of correctly labeled negative results over all negative labeled results.

$$n := \frac{TN}{TN + FN} \tag{3.3}$$

It is a measure of a classifier's reproducibility of the negative results.

Specificity s (a.k.a. true negative rate) is the percentage of correctly labeled negative results over all negative cases.

$$s := \frac{TN}{FP + TN} \tag{3.4}$$

It is a measure of a classifier's ability to separate out the negative cases.

Fall-out f (a.k.a. false positive rate, false alarm rate) is the percentage of incorrectly labeled positive results over all negative cases.

$$f := \frac{FP}{FP + TN} \tag{3.5}$$

It is a measure of a classifier's tendency to include non-relevant results.

Accuracy a is the percentage of correctly labeled results over all results.

$$a := \frac{TP + TN}{TP + TN + FP + FN} \tag{3.6}$$

It is a measure of a classifier's veracity (conformity) to the classification condition.

3.2.3 Combined Measures

Matthew's Correlation Coefficient - MCC

Matthew's correlation coefficient is a balanced measure of a test's results.

$$MCC := \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.7}$$

This coefficient expresses the quality of binary classifications. Its value ranges from -1 to $+1$: 1 represents the perfect classification, 0 the average random prediction result, and -1 the inverse classification. If any of the denominator's four sums is zero, the denominator is set to one; This results in a coefficient of 0. It is a more stable measure than accuracy for cases where the class distribution is imbalanced.

F-measure

The F-measure F_β is the harmonic mean between precision and recall, where β is a parameter for the relative importance of precision over recall.

$$F_\beta := (1 + \beta^2)\frac{p \cdot r}{\beta^2 p + r} \tag{3.8}$$

The balanced F-measure ($\beta = 1$, a.k.a. F-score) can be simplified to:

$$F_1 = 2\frac{p \cdot r}{p + r} \tag{3.9}$$

Note that the F-measure does not take the TN rate into account and therefore MCC usually is preferable to assess the performance of a binary classifier. However, in many IE tasks, the TN rate cannot be established (or is a "virtual constant"[3]) and therefore the F-measure is the more frequently applied IE measure.

Receiver Operator Characteristic - ROC

A receiver operator characteristic curve measures the quality of a ranked result set by opposing the true and false positive rates (recall r and fall-out f, respectively). Compared to F-measure, MCC score, and accuracy, the $A(ROC)$ can evaluate ranked (ordered) results. The area under the ROC curve (AUC ROC) is defined as:

$$A(ROC) := \int_0^1 f(r)dr \approx \sum_i^n f(r_i)\triangle r(r_i) \tag{3.10}$$

To calculate an approximation of the $A(ROC)$, the following function is used:

$$A(ROC) = \sum_i^n \frac{f(r_i) + f(r_{i-1})}{2} \cdot (r_i - r_{i-1}) \wedge r_0 = 0, f(0) = 0 \tag{3.11}$$

[3] I.e., if the number of negative cases is significantly larger than the positive cases and the result sizes are very small, the TN rate is nearly constant.

Where r_i is the recall, and f the fall-out measured at recall r_i. For each ranked result, the accumulated recall and fall-out up to that result position are calculated. Then these r, f pairs are ordered by increasing recall, then fall-out. These pairs are used to plot the ROC curve and calculate the AUC.

Precision/Recall - PR

Just as ROC curves, PR curves evaluate ranked (ordered) results. For each ranked result, the accumulated recall and precision up to that result position is calculated. Then these r, p pairs are ordered by increasing recall, then decreasing precision. The r, p pairs can be used to both plot the PR curve and calculate the AUC A. Note that a plot of the curve provides visual information about a classifier's tradeoffs.

The AUC PR measures the quality of a ranked result set. The area A under the precision p - recall r curve of n ranked results is defined as:

$$A(PR) := \int_0^1 p(r)dr \approx \sum_i^n p(i)\triangle r(i) \tag{3.12}$$

The approximation of $A(PR)$ in Equation 3.12 is also known as the *average precision* (AP, next subsection). To calculate the most exact approximation of the area, the following function is applied:

$$A(PR) = \sum_i^n \frac{p(r_i) + p(r_{i-1})}{2} \cdot (r_i - r_{i-1}) \wedge r_0 = 0, p(0) = 1 \tag{3.13}$$

Where n is the number of results ordered by increasing recall, and r_i is the recall at result i. Setting $p(0) = 1$ (instead of choosing $p(0) = 0$) reduces the impact of having a hit in the first position as opposed to only having the first hit in one of the follow-up positions. Assuming only one result is relevant, if $p(0) = 1$ and the first result is a hit, AUC PR is 1.0, if it is the second, AUC PR is 0.75, and 0.5 if only the fourth result is a hit. If $p(0) = 0$, the same hits add 0.5, 0.25, and 0.125 to the AUC. Not only does the AUC never reach 1.0, but having the first position correct adds double as much area compared to having the first hit in the second position.

Average Precision - AP

AP is based on the PR integral, just as AUC PR. However, the area under the approximated (see Equation 3.12) curve is calculated as:

$$A(PR) \approx AP := \sum_{i}^{n} p(r_i) \cdot (r_i - r_{i-1}) \wedge r_0 = 0 \tag{3.14}$$

Where n is the maximum number of ranked annotations on any article. Due to the curve approximation, the weight of a hit in the first position is stronger than for AUC PR values, but AP, too, can reach 1.0 if all relevant results are in the top positions.

Threshold Average Precision - TAP

The threshold average precision measures the quality of a ranked result set up to a cutoff of k wrong predictions. TAP is defined as the average precision up to a *sentinel* record at position n, where a total of k false positive classification have been encountered or the end of the list has been reached (Carroll et al. 2010). Thresholded average precision is calculated by summing the precisions p_i of all hits (TP) m reported up to the sentinel[4] at n plus the precision at the sentinel itself (p_n), and then dividing this sum by the total number of relevant annotations (i.e., $TP + FN$) in the set:

$$TAP_k(R) = \frac{\sum_{i}^{m}(p_i) + p_n}{TP + FN + 1} \tag{3.15}$$

Where p_i is the ith precision value measured at one of the m relevant records before the sentinel, and the denominator represents all true cases plus a corrective term to account for adding p_n.

In essence, TAP is the same as AP, but with a cutoff applied and the sentinel's precision added (because the sentinel does not have to be a hit). TAP does not evaluate any results after the sentinel, and the cutoff k represents a user's tolerance towards errors.

3.2.4 Agreement Measures

The following two metrics (JPA and the Kappa coefficient) are used to evaluate the degree of agreement among of categorical annotations, also known as the *inter-annotator*

[4] therefore, either the last or a FP result

agreement (IAA). They are applied in this work to measure the quality of the expert-curated "gold standard" sets. These measures help estimate the best performance an automated system can achieve if it can be assumed that a machine learning system cannot do better than the quality of data it is learning from.

Joint Probability of Agreement - JPA

The joint probability of agreement J is defined as the percentage of times two annotators A and B made the same categorical annotation i on n instances:

$$J = \frac{\sum_i^n |A_i \cup B_i|}{n} \tag{3.16}$$

When the number of categories being used is small, the likelihood for the annotators to agree by chance increases dramatically. This chance agreement is counter-acted by their propensity for intrinsic agreement (i.e., agreements not based on chance). However, JPA will remain high, because it does not take this chance agreement into account and always assumes intrinsic agreement. An advantage of the JPA is that it is a comparable measure to the F-score, while the Kappa coefficient (below) is not.

Kappa Agreement

The Kappa coefficient κ is an agreement measure for categorical classifications, just as JPA, but takes agreement by chance into account.

$$\kappa = \frac{p(\omega) - p(\epsilon)}{1 - p(\epsilon)} \tag{3.17}$$

Cohen's Kappa is used for comparing two annotators, Fleiss' Kappa is an extension of Cohen's Kappa to compare multiple annotators. In this work, only Cohen's equations for computing the agreement probability were applied. Cohen's $p(\omega)$ is the observed probability of agreement, and the same as the JPA:

$$J = p(\omega) = \frac{\sum_i^n |A_i \cup B_i|}{n} \tag{3.18}$$

$p(\epsilon)$ is the expected probability of agreement, or the "chance agreement", defined as the sum of chance agreements over all category labels L:

$$p(\epsilon) = \sum_{j}^{L} \frac{|A_j| \cdot |B_j|}{n^2} \tag{3.19}$$

Where $|A_j|$ and $|B_j|$ are number of times annotator A and B, respectively, assigned label j on any of the n instances. This term corrects the JPA by the probability of the annotators choosing the same labels by chance. The Kappa coefficient is therefore more robust than JPA. However, Kappa agreement factors chance as a constant and has no measure for intrinsic agreement. Therefore, it can be considered a rather conservative measure.

3.2.5 Task Metrics

For the article classification task, the classification performance was measured with Matthew's correlation coefficient (MCC). The ranking performance in the same task was established with the AUC PR mesure. The PR curves themselves were plotted with the evaluation software. In addition, sensitivity, specificity, and accuracy were measured.

The classification performance of systems participating in the protein normalization, interaction detection, and method extraction task was established with the F_1-measure (F-score). In addition, precision and recall were measured. The ranking performance was measured with the Average Precision (AP) metric, which puts more emphasis on a hit in the first position. However, there is no meaningful single ranked results list for all annotations across all articles in these tasks. Therefore, p, r pairs for the AP function are generated at each rank across all articles. Precision and recall are calculated from the total TP, FN and FP counts over all articles up to each rank in the micro-averaged scenario or by averaging the precision, recall values at each rank across all articles in the macro-averaged scenario (see Section 3.3.3 for these scenarios). The AP curves were plotted with the evaluation software. Finally, to measure the combined impact of classification and ranking performance, a new measure, the FAP-score is introduced in this work.

F-measured Average Precision

The F-measured Average Precision (FAP) score presented in this work is the harmonic mean between the F-measure and AP score. The FAP-score is defined as:

$$FAP_\beta := (1+\beta^2)\frac{F_\beta \cdot AP}{\beta^2 F_\beta + AP} \tag{3.20}$$

Just as when weighting the contribution of precision to the F-measure, the β parameter proportionally increments the importance of the F-measure in the FAP-score with increasing β. The F-measure represents the accuracy of the results, i.e., it ensures high precision, while the AP tends to increase recall. Therefore, the same β parameter can be propagated to the F_β function itself. For the balanced situation ($\beta = 1$), the FAP-score is simplified to:

$$FAP_1 := 2\frac{F_1 \cdot AP}{F_1 + AP} \tag{3.21}$$

The FAP score is a metric that dynamically penalizes long result lists without relevant hits due to the F-measure and at the same time incorporates ranking performance with the AP term. A perfect FAP-score (1.0) is only achieved with a perfect result set (i.e., all relevant results and no irrelevant results are reported), and it thus is an objective optimization function that identifies the best possible ranked result list (in terms of AP) that at the same time has an optimally limited result set size (in terms of F-measure). As both F-measure and AP themselves are rates, the harmonic mean needs to be applied.

3.3 Challenge Evaluations

3.3.1 Result File Format

The result file format BioCreative participants were asked to submit are tab-separated value files (.tsv) for the BioCreative II.5 and III challenges, and for the article classification task of BioCreative II. Each line in the .tsv file represents an annotation result. The values (rows) needed to contain:

- the article identifier (the PubMed ID for BioCreative II and III; or the article's DOI for BioCreative II.5 on the training set and a randomized article ID for test set),
- the protein identifier(s) as UniProt accession number(s) for the protein normalization (one accession) and PPI interaction pair (a pair of accessions) tasks, the

PSI-MI term ID for the method extraction task, or a Boolean value (true, false) for the article classification task, and

- a normalized confidence value for the prediction in the range $(0, 1]$.

Optionally, if the confidence values do not fully reflect the desired result order, the rank of each result can be given explicitly, in addition to a confidence value. Ranks have to be positive, unique integers (with respect to the entire result set in the article classification task, and with respect to a single article's annotations in the other tasks). If given, ranks were preferred over confidence values for the ranking evaluations.

For BioCreative II, an XML format was used to represent each result of the protein normalization, interaction detection and method extraction tasks. Each article's result had to be represented by an `ENTITY` element. Each `ENTITY` had the following child nodes: `PPI_SUB_TASK_ID` with the task ID (interaction detection or method extraction), the team ID in `TEAM_ID`, the run ID in `RUN_NR`, and the PubMed ID in `PMID`. For the protein normalization and interaction detection task, all article-specific `ENTITY` results had to be grouped into one or more `INTERACTION_PAIR` elements that contained three leaf nodes: `INTERACTOR_1` and `INTERACTOR_2` with the UniProt accessions for each interacting protein and a `RANK` element as a unique rank with respect to the result `ENTITY`. For the method extraction task, one or more `INT_DET_METHOD` elements had to be attached to the `ENTITY` nodes, containing a `RANK` element as described and a `INT_DET_METHOD_ID` containing the PSI-MI ID of the method.

It should be noted that hardly any team produced this XML format correctly, however. Results for a single article were spread across multiple `ENTITY` nodes with non-unique ranks (e.g., all annotations in a single article were given a rank of 1). Often, the XML itself was illegal, such as missing closing tags, and a robust parser for the bad XML had to be developed. Most teams did not report UniProt accessions, but the (unstable) IDs (UniProt IDs are only valid for one specific release). This meant that all such UniProt IDs had to be mapped to the correct accessions using the UniProt database. Similarly, some teams did not report PSI-MI IDs but names of PSI-MI terms. In this case, too, the names had to be mapped back to the actual IDs. The XML format was abandoned for all later challenges and a rigid file content check implemented in the BioCreative evaluation tool (see Section 3.3.2).

3.3.2 Evaluation Software

All evaluations were made using a Python library developed specifically for this task (the "evaluation software"). It checks the correctness of the tabulated input (for both result files and the gold standard), and is able to perform the evaluations for all scenarios described in this work, including a homonym ortholog mapping (not described in this work) and the organism filtering evaluations (see Section 3.3.3, Organism Filtering Evaluation). The tool reports the number of annotated documents and total annotations, the number of true and false positives/negatives, and the MCC/accuracy as well as the AUC PR (article classification task) or the micro- as well as macro-averaged F-measure, AP, and FAP-score (all other tasks). For the macro-averaged scores, the standard deviations of precision, recall, and F-measure are calculated. The AUC PR and the (micro- and macro-averaged) AP curves can be plotted by the tool, which makes use of the `matplotlib` Python plotting library. After extracting the initial values, all further processing and evaluation, including the Kappa inter-annotator agreement (see Section 3.2.4) calculations, was completed in the R statistics environment, using the `ggplot` package for generating result plots. The overlap evaluation (authors-curator-system) Venn diagrams are plotted with the Google Chart API[5], and a simple Ruby script was written to interact with it.

3.3.3 Evaluation Scenarios

Macro- vs. Micro-Averaged Evaluations

For the protein normalization, interaction detection and method extraction tasks, results are reported per article. In the micro-averaging procedure, first the TP, FP, TN and FN counts from all articles are established and then the metrics are taken using those counts. In the case of macro-averaging, the counts are made for each individual article and the scores for each article established. Then, all the individual article scores are averaged by taking the mean to produce the final results. This procedure has the advantage that distortions due large imbalances between the number of annotations made per article – a phenomena inherent to the BioCreative PPI corpora, where an article might be annotated with a just a single PPI or up to 30 – are reduced. For the two PPI corpora tasks (protein normalization and interaction detection), macro-averaging produces a more generalized system performance value than micro-averaged calculations would. On

[5] http://code.google.com/apis/chart/

the other hand, for the method extraction task, the number of annotated methods per article is far less variable, and the micro-averaged results are reported. It is worth clarifying that micro-averaging, despite its name, does not include taking any averages at all.

Document-Centric Evaluation

By default, all metrics presented here are calculated from the complete results set using all articles. In this work as well as the official BioCreative results presented in the respective publications, another evaluation strategy is to calculate the performance measures only for articles for which the system produced results. In other words, document-centric evaluation excludes articles for which the classifier (results) produced no annotations at all. This evaluation approach was applied to the protein normalization and interaction detection tasks. The document evaluation scenario evaluates a classifier that only annotates those article for which it has sufficient confidence in the quality of its annotations. This document-centric approach is important both in high-precision annotation scenarios or to evaluate the performance a (human) user would perceive under the condition that the BioNLP system produced any results at all.

FAP-score-based Cutoffs

Some protein normalization and interaction detection submissions during BioCreative II.5 maximized their AP scores by producing very long result lists. To estimate the maximum F-score these runs could have achieved, the confidence value cutoff that maximizes their FAP-score is determined. We scanned their *training set* FAP-scores in 0.01 increments to find the confidence score cutoff that has the highest FAP-score. I.e., this cutoff is the confidence score after which all further results with lower confidence are ignored (dropped from the result list). Then, this training set confidence score cutoff is used on their test set results to establish the highest FAP-score the run might have achieved. In other words, the FAP-score based cutoffs establish the results of these runs had they been optimized for both F-measure and AP, instead of AP only.

Organism Filtering Evaluation

Species (organism) identification is a hard task for automatic annotation because authors often do not mention the species, descriptions of experiments carried out with proteins from multiple species are common in the PPI literature, or the protein(s) of a mentioned

species (especially human) might not have been used by the authors. As a practical solution, in addition to the raw system results, we applied a filter to estimate how good the automated results are if filtered by the organism information, as could be provided by the author or curator.

First, the relevant organism are established by mapping the gold standard proteins to their taxonomic IDs, using the UniProt release 15.0 data. Then, all proteins annotated by the system under evaluation are mapped to their species, too. All annotated proteins that do not map to relevant (gold-standard) organisms are removed. Only the remaining proteins are used for calculating the evaluation results.

3.4 Linguistic Annotation Types

From a language processing perspective, the types of annotations that can be made on a document have been categorized into five groups of document- and sentence-level annotations.

3.4.1 Document Annotations

Document annotations are slightly different from all other annotation types and not strictly linguistic, as they are assigned to an entire piece of text, for example tagging documents for classification. A BioCreative-specific example is the annotation of a list of genes mentioned in a text by their DB IDs without mapping these annotations to the actual locations of the mentions in the text, or a ranked list of PPIs in a publication without mapping them to the actual passages in the text that describe the interaction. All annotations made in the context of the BioCreative PPI tasks presented in this work pertain to this annotation category. However, to generate these data, systems commonly will produce several - if not all - of the other annotation types in the process of identifying the most relevant document annotations.

3.4.2 Lexical Annotations

Lexical annotations are annotations of "tokens" (e.g., words, numbers, symbols, syllables, etc.) with their lexical (words, letters, numbers, alphanumerics, punctuation, symbols, spaces, etc.) or grammatical class - termed "Part-of-Speech" (PoS): one of verb, noun, pronoun, adjective, adverb, preposition, conjunction, or interjection. This includes the word *morphology*, such as the use of upper- and lower-case letters. Depending on the

particular lexical tagging system, these eight PoS super-classes are divided into several subcategories. Generating lexical annotations is a virtually ubiquitous pre-processing step of BioNLP systems. Lexical annotations are sentence-level annotations.

3.4.3 Syntactic Annotations

Syntactic annotations tag the linguistic structure of a sentence. This can range from the shallow syntactic structure (annotating the boundaries of sentences, as well as the phrases and clauses inside the sentence), to the full grammatical relationships of a sentence' phrases (known as the syntax tree). Due to the fact that BioCreative puts its emphasis on the biological and not the linguistic content, syntactic annotations are not present in any of the BioCreative corpora, but especially for the interaction detection tasks, many systems generate these annotations.

3.4.4 Semantic Annotations

Semantic annotations are context-independent annotations on text passages. The simplest kind of such annotations would be to tag, for example, the mention of a gene as a "gene name" concept, a form of Named Entity Recognition (NER). From a linguistic perspective, these semantic meanings determine lexical and syntactic annotations. This annotation class is the basis for the gene mention task of BioCreative I and II. However, for the tasks described in this work, systems were not required to map the protein normalizations to their text passages.

3.4.5 Discourse Annotations

Discourse annotations finally are context-dependent annotations on text passages, i.e., in the context of an entire piece (discourse) of text, and includes (outside) "world knowledge". This involves resolving expressions referring to another object somewhere in the text, identifying spatial or temporal relations, extracting the theme ("rheme") and context of expressions, etc.. Gene normalization (GN; tasks during BioCreative I, II, and III) annotations, while provided at document-level, are examples of context-dependent annotations. In many cases, the correct normalization (i.e., mapping of a gene name to a database identifier) can only be made by taking into account the entire context of a document, such as species mentions, genomic locations, or other relevant pointers present throughout the discourse of a document.

3.5 BioCreative Meta-Server

3.5.1 Libraries and Architecture

The BioCreative Meta-Server is written in Python (back-end) and Javascript (front-end) communicating via AJAX (Asynchronous Javascript And XML). Upon a user request, the article is immediately presented to the user, while the meta-server fires annotation requests to annotation servers. Therefore, the AJAX API is used to update[6] the article's annotations as they are received from the servers. The web front-end itself is developed on top of the Django web development framework and is served by a LigHTTPD web server. The MySQL database is accessed by the MySQLdb Python DB broker.

The meta-service (back-end) connecting to the annotation servers is completely written in house in Python and the design will be detailed in the results. The web-service API for requesting annotations from the servers is implemented via the XML-RPC protocol. On the meta-server, it is implemented using the Python `xmlrpc` standard library. Result validation (for the annotations received from the annotation servers) is a two-step process. First, the meta-server verifies the correct syntactic structure of identifiers and meta-data, and then DB-specific web services are contacted to ensure the reported identifiers exist in the actual databases (UniProt for proteins, Entrez for genes). Once an identifier and its meta-data (e.g., the gene names) has been retrieved, it is stored locally in a separate database table. This reduces the time required by the verification process, as each identifier is requested only once, no matter how many annotation servers report it. Text data (PubMed abstracts) are imported to the database by connecting to the NCBI eUtils web service. The full system architecture is shown in Figure 3.1. In addition, for the challenges, the Django user management module was activated and an e-mail-based account validation system was implemented. These features were required so participants could log on and manage their annotation servers.

[6] using polling

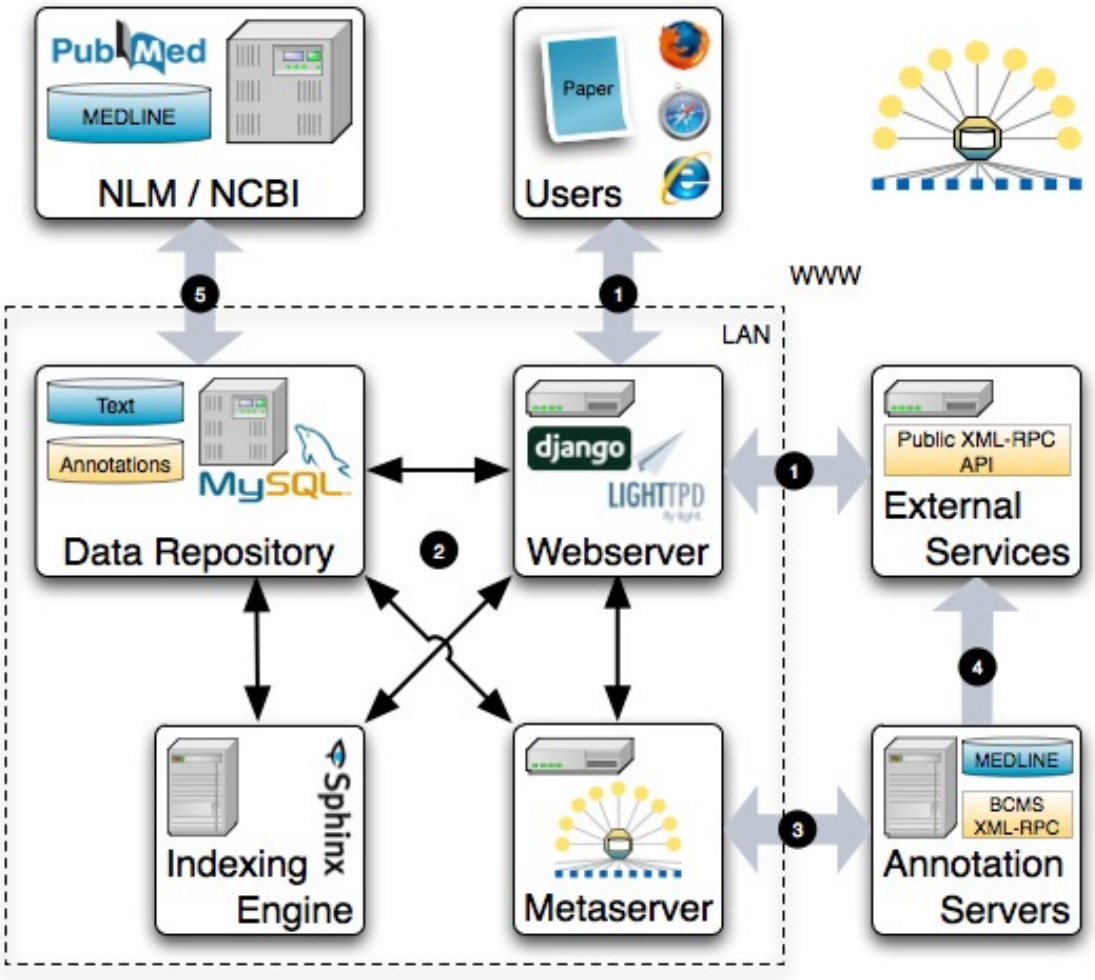

Figure 3.1: An overview of the BioCreative Meta-Server. The BCMS consists of four components: A data repository for storing BioNLP annotations; An index to make queries against the deposited data; A web server for serving the results to users and any other services; And the actual "meta-server" sending annotation requests to BioNLP pipelines running as web services (Annotation Servers) and processing responses from those services. Commonly, a (human) user or an external service will initiate a request for annotating a MEDLINE abstract (1) and the web server will check the if any annotations for the abstract already exists in the repository (2). If they do, these results are served. If not, the meta-server initiates a request for the abstract (3) to the annotation servers, and returns a collation of the results to the user/external service after they have been validated. As all BioNLP pipelines are running as separate services, an external service can communicate with those pipelines directly, too (4). New abstracts can be imported/added from PubMed (5).

CHAPTER 4

Results

Challenge	train ⊕	train ⊖	test ⊕	test ⊖
BioCreative II	3,536	1,959	338	339
BioCreative II.5	61	534	63	532
BioCreative III	1,822	4,458	910	5,090
Total	5,419	6,951	1,311	5,961

Table 4.1: BioCreative PPI **article corpora** sizes. **train**: training set; **test**: test set; ⊕: positive (relevant) articles; ⊖: negative (irrelevant) articles. Total: 19,642 documents.

4.1 The BioCreative PPI Corpora

A substantial contribution of the BioCreative PPI efforts is to produce datasets with high-quality annotations that can be used as common basis for the evaluation of PPI extraction systems. In the community jargon, a dataset of text and annotations curated by expert annotators is commonly referred to as a "gold standard".

For each challenge, a unique corpus was created in cooperation with professional curators from IntAct, MINT, BioGRID, as well as specifically trained biologists. These manual annotations were later corrected by the BioCreative organizers during the course of the challenge whenever errors were encountered (and after requesting curator feedback on a proposed update[1]). Multiple reviews and annotations by curators, organizers, and even participants ensure the final corpora stemming from these challenges represent high-quality, professional PPI annotations. Overall, the BioCreative PPI corpora span nearly twenty thousand abstracts classified as PPI-relevant or not and 3,632 full-text articles annotated with experimental methods or interaction pairs, plus another 1,066 ("negative") full-text articles without PPI information (used in BioCreative II.5). The methods for assembling the corpora have been described in Section 3.1. For all challenges, many different publishers provided the organizers with permissions to redistribute the PDF, XML, or HTML files of publications.

4.1.1 BioCreative PPI Article Corpus

For the PPI article classification tasks of BioCreative II-III, professional DB bio-curators reviewed 19,642 abstracts and judged their relevance for PPI curation following their in-house curation protocols. The exact distribution of relevant and non-relevant articles is shown in Table 4.1.

[1]It should be noted that updates were required for no more than approximately 1-2% of the PPI annotations.

For BioCreative II.5, instead of providing only abstracts, the full-text articles were made available to the participants. This was possible due to Elsevier, the publisher behind FEBS Letters, giving permission to the BioCreative organizers to distribute the entire article data of FEBS Letters for the years 2007 and 2008. In addition, Elsevier supplied the organizers with the XML files for all articles used in the challenge, thereby removing the PDF/HTML extraction issue. The BC II and III articles were based on a mixed selection of journals. The balanced test set of BioCreative II is expected to result in higher ranking scores (AUC PR) than the other two sets sets.

For the BioCreative III classifications, an inter-annotator agreement (IAA) study was performed. Scores between MINT, BioGRID, and the expert curators were measured. There is a very strong agreement on labels between the two database curators (96% JPA, $\kappa = 0.85$). BioGrid and MINT follow (similar) in-house protocols for labeling the abstracts while the expert annotators used a protocol designed specifically for the challenge. This resulted in a better agreement between database curators than between the expert annotators and the curators (for MINT vs. expert, 92% JPA & $\kappa = 0.69$; for BioGrid vs. expert, 91% JPA & $\kappa = 0.69$). There was a three-way agreement on the binary labels between all three groups (MINT, BioGrid, and expert annotators) on 85.5% (JPA) of all articles. These agreements are comparable to the best accuracy measured in the task (89%, see Table 4.7).

4.1.2 BioCreative PPI Pair Corpus

For the PPI protein normalization and interaction detection tasks, the annotations were created by IntAct and MINT curators. The two tasks were only part of the II and II.5 challenges. Altogether, this effort produced a corpus of over one thousand full-text articles with about five thousand PPI pairs, annotated by their UniProt accessions (see Table 4.2). All accessions used to annotate the proteins (and pairs) are part of UniProt major release 15 or newer[2]. Slightly more than 10% of all interactions in the data sets are cross-species, while the large majority are pairs where both interaction partners are from the same species.

The 367 BioCreative II test set articles have a total of 1213 unique UniProt accessions annotated 1393 times (protein normalization task), and forming 1409 interaction pairs (interaction detection task). Of these 1213 accessions, 778 are SwissProt accessions, and 435 are TrEMBL accessions. 242 out of the 346 articles have only SwissProt annotations,

[2]because UniProt never replaces accessions

Challenge	**train**	pairs	appa	**test**	pairs	appa
BioCreative II	628	3179	5.0	346	1332	3.8
BioCreative II.5	61	236	3.9	61	216	3.5
Total	689	3415	4.9	407	1548	3.9

Table 4.2: BioCreative PPI **pair corpora** sizes (protein normalization and interaction detection tasks). **train**: training set articles; **test**: test set articles; **pairs**: total PPI pairs annotated; **appa**: average PPI pairs per article; Total: 1,096 documents annotated with 4,963 PPI pairs.

Task	Intra-DB JPA	Intra-DB κ	Inter-DB JPA	Inter-DB κ
Protein normalization	95%	0.84	86%	0.75
Interaction detection	87%	0.74	76%	0.36

Table 4.3: BioCreative II.5 pair corpus **IAA**. JPA measurments and κ coefficients comparing annotation agreement between curators from the same DB (Intra-DB) or two different DBs (Inter-DB).

leaving 104 that have mixed SwissProt and TrEMBL annotations. In the BioCreative II corpus, 1130 UniProt interaction pairs are based on SwissProt accessions only, while only 202 pairs contain at least one TrEMBL identifier. The 242 SwissProt only articles have 883 protein normalizations forming 834 interaction pairs.

The 61 BioCreative II.5 test set articles have a total of 246 unique UniProt accessions annotated 252 times (protein normalization task), and forming 216 interaction pairs (interaction detection task). Of these 246 accessions, 229 are SwissProt accessions, and 17 are TrEMBL accessions. 46 out of the 61 articles have only SwissProt annotations, leaving 15 that have mixed SwissProt and TrEMBL annotations. In the BioCreative II.5 corpus, 192 UniProt interaction pairs are based on SwissProt accessions only, while only 24 pairs contain at least one TrEMBL identifier. The 46 SwissProt only articles have 194 protein normalizations forming 164 interaction pairs.

The BioCreative II.5 articles that had been annotated by MINT curators were in parts annotated twice, one set by another MINT curator (20/61 articles, intra-DB agreement) and by an IntAct curator (another 21/61 articles, inter-DB agreement). Agreements for the protein normalization and for the interaction pairs were measured as shown in Table 4.3.

One of the major difficulties for protein normalization is associating the proteins to the correct organisms. Figure 4.1 shows the organism distribution of the BioCreative II.5 pair corpus. The corpora do not aim to create a perfect balance. They only a reflect a partially similar distribution. This situation mimics the real-world setting where any

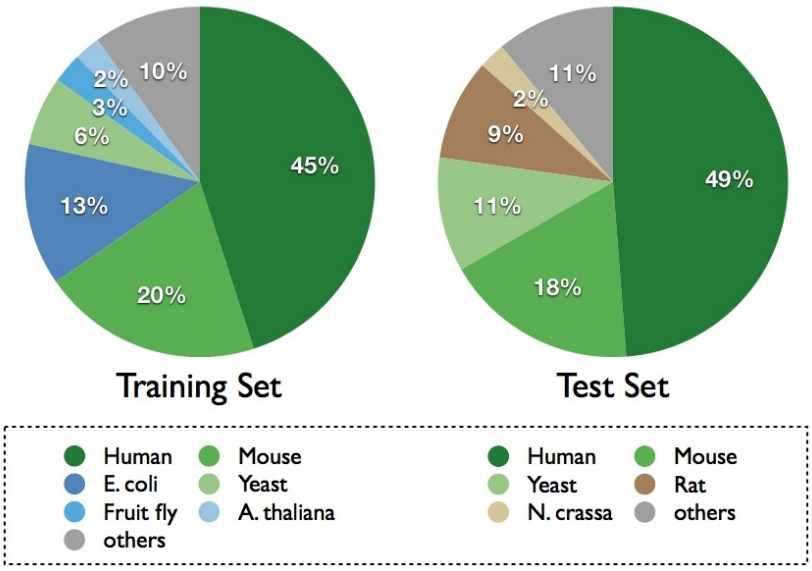

Figure 4.1: Distribution of proteins by organism, grouped by the five most frequent organisms annotated on the BioCreative II.5 pair corpus.

Challenge	**train**	meth	ampa	**test**	meth	ampa
BioCreative II	740	1640	2.2	368	874	2.4
BioCreative III	2003	4348	2.2	223	527	2.4
Total	2699	5891	2.2	590	1401	2.4

Table 4.4: BioCreative PPI **method corpora** sizes. **train**: training set articles; **test**: test set articles; **meth**: total experimental methods annotated; **ampa**: average methods per article; Total: 3,289 documents annotated with 5,749 methods. Note that 44 articles overlap between the BioCreative II and III training set.

PPI article that exists should be annotated, independent of organism focus.

4.1.3 BioCreative PPI Method Corpus

The PPI method extraction tasks formed part of the BioCreative II and III challenges, and the annotations were made by IntAct (BC II only), MINT (BC II and III), and BioGRID (BC III only) curators. Both together produce a corpus of 3,289 full-text articles with over five thousand annotated experimental methods as PSI-MI term IDs (see Table 4.4). In total, 115 PSI-MI terms describe experimental methods that can be used to validate a PPI.

For BioCreative II, in total 70 (out of the 115) methods are annotated on the training set and 60 on the test set. However, 11 methods in the test set have no examples in

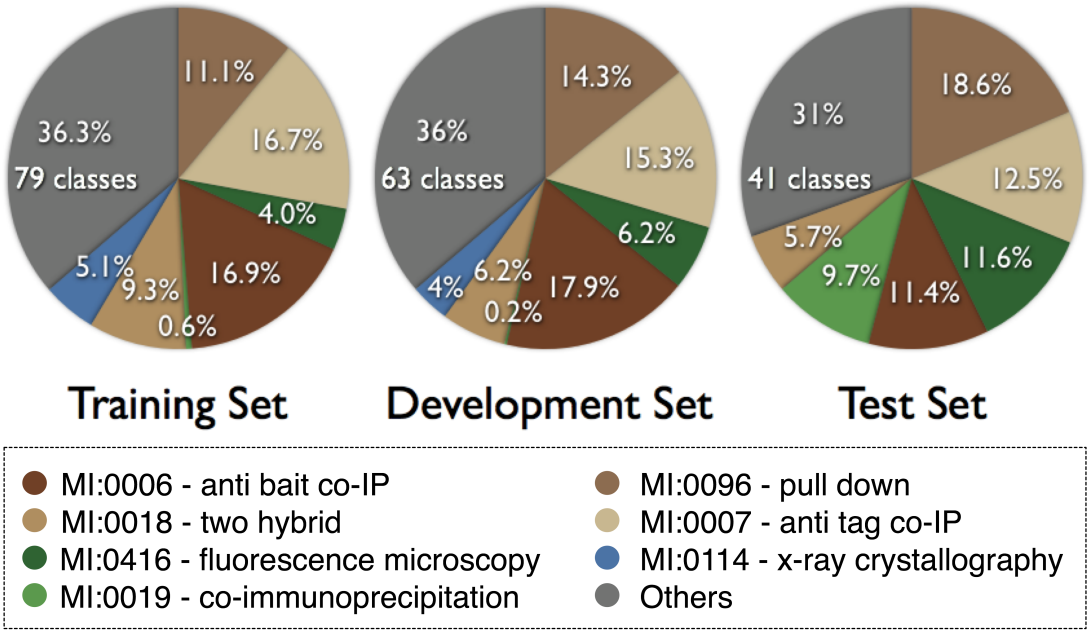

Figure 4.2: Distribution of the most frequent experimental method annotations in the three sub-sets of the BioCreative III method corpus.

the training set, while 21 examples in the training set are not part of the test set, and therefore only 49 methods are annotated in common. This makes it very difficult for supervised approaches learning exclusively from the training examples to perform well on the test set.

To lessen the issue of methods only found in the test set and provide more examples per method, during BioCreative III, participants were supplied with examples for nearly all methods that appeared in the test set. Only one experimental method annotated in the test set has no example in the training set. To achieve full coverage, in addition to the articles listed in Table 4.4, an additional set of 587 articles was handed out to participants of the BioCreative III challenge shortly before the test phase, as a "development set". With this development set, the total coverage of methods annotated increases from 86 to 97. The training plus development set provide examples for 46 (out of 47) experimental methods annotated on the BC III test set. As can be seen from the pie charts in Figure 4.2, a large majority of articles are annotated with a small number of frequently used experimental methods.

4.2 Challenge Settings

The instructions and announcements regarding the challenges - both offline and online - were published on the BioCreative web page[3], included in the downloads[4], and published as news feeds[5] and via a dedicated mailing list[6].

4.2.1 Settings for the Offline Evaluations

For the offline evaluations, participants could download all articles and the gold standard annotations for the training sets. Regarding the tasks of BioCreative III and the article classification task of BioCreative II, additional sets of articles and annotations were released just weeks before the test phases as "development sets". These additional releases were made to provide participants with class distributions that best reflect the distributions found in the test sets.

From BioCreative II.5 on, participants were supplied with an installer for the evaluation library as a download from the BioCreative web site[7]. This allowed participants to both verify the syntactic correctness of their results and to evaluate the performance of their system on the training data using the same procedure applied during the evaluation phase. During BioCreative II.5 (only), participants were asked to provide their final training set runs prior to receiving access to the test set. This requirement allowed the organizers to verify the result data produced by the teams and compare their training and test set results.

Upon entering the test phases, participants were provided with a link to the download of the articles, but no annotations were supplied with them. The day this link was sent was used as the reference for a one week time limit teams had to obey when reporting results. Within that week, teams were asked to produce up to three (BC II) or five (BC II.5 and III) result sets for each task they wished to participate in. One day before the time frame expired, teams were notified about the impending expiration date via e-mail. The final result sets could be uploaded to a FTP server provided by the organizers or sent directly via e-mail.

[3] http://www.biocreative.org/ - see sections Events and Tasks
[4] http://www.biocreative.org/resources/corpora/ - requires registration
[5] http://www.biocreative.org/feeds/all/
[6] biocreative-participant@lists.sourceforge.net
[7] `http://www.biocreative.org/resources/biocreative-ii5/evaluation-library/`

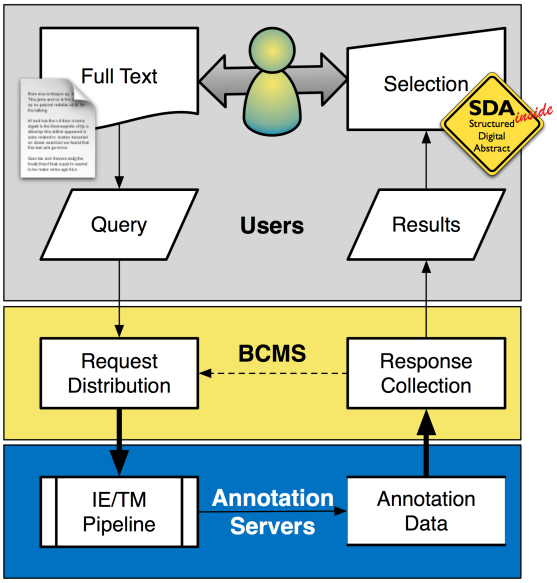

Figure 4.3: Diagram of the online evaluation scenario. An author or database curator submits an annotation query for a PPI manuscript to the BCMS. The BCMS distributes the request to the annotation servers and waits for their results. The servers run their BioNLP pipelines on the text and extract PPI information. The analysis results from the annotation servers are collected and validated by the BCMS. The results are then presented as integrated data to the user. The user would select relevant results and adds missing results, thereby producing a structured dataset describing the content of the manuscript, such as a Structured Digital Abstract (SDA). Dotted arrow: In case of communication failures (or, outside the challenges, validation errors), the annotation request to a server is repeated.

4.2.2 The Online Evaluation Scenario

The BioCreative II.5 and III challenges were carried out online using an modified version of the BioCreative Meta-Server described in the next section. Contrary to the initial BCMS version created after BioCreative II, this extended BCMS allows participants to take direct control of and monitor the communication between their systems and the BCMS. This creates an environment simulating a realistic, distributed web service model inter-operating with the participant systems. The model online scenario including a human consumer of the systems is depicted in Figure 4.3.

4.2.3 Settings for the Online Evaluations

To allow participants to control the interaction with the BCMS during the challenges, several modifications and extensions to the initial BCMS design were made. A management interface for each team was added, where teams could register each annotation server with its URL, and select the tasks for which the server reports results. The

interface allows to manage the server, inspect logs, and view the current annotation states for each server. The public interface to the BCMS (described in Section 4.4) was deactivated.

The team administrator can load the articles into a queue for each annotation server, and instruct the BCMS to send the articles one by one to the registered annotation severs. The BCMS expects to receive a result for each request (i.e., article) within ten minutes, as a XML-RPC structure reporting the UniProt accessions (protein normalization) or pairs (interaction detection), the PSI-MI IDs (method extraction) or a Boolean value (article classification task) together with a confidence score. The interface allows to halt the annotation process at any time[8].

In the case that (web/protocol) communication errors with the annotation server occur during the request cycle of an article or the corresponding result is invalid, the BCMS tries to repeat the request five times before moving on to the next article. Furthermore, each time such an error occurs, an error analysis process determines the most likely reason for the error. The outcome of this analysis is logged together with the error itself. An e-mail notification system for errors and successful runs can be configured by the teams. An interface to manage and inspect each server's error logs is available through the management interface. Errors might range from connection problems[9] and communication problems [10] to result errors [11]. At any time during the training phase, teams can request the performance results and/or re-run the entire annotation process again.

Before teams were allowed to move into the test phase, they had to successfully complete the entire run on the training set. During the test phase, re-queueing the article set was disabled. A hard limit of 250 normalization/pair results per article was enforced, because more results would be of no other utility than increasing the AP score. If communication or connection errors occur during the test phase, the process is halted (i.e., automatic retries are not triggered) and teams can decide to move on to the next article, resulting in an empty annotation set for the article being processed when the error occurred. Alternatively, the annotation server administrator may instruct the BCMS to retry the request-response cycle for that article[12]. However, after the fifth failed attempt,

[8] Gracefully, i.e., waiting for the latest request to finalize before shutting down the process, or hard, as emergency fallback.

[9] e.g., the server not running, the service port being unreachable, the URL being incorrect, etc.

[10] invalid XML, broken packages, etc.

[11] missing or malformed values, incorrect string encoding, illegal offset values, non-existing database identifiers, etc.

[12] but only for communication and connection errors

Challenge	AC	PN	ID	ME	Total
BC II	19	16	16	2	26
BC II.5	8	10	9	n/a	15
BC III	10	na/	n/a	8	11
Total	37	26	25	10	52

Table 4.5: Number of teams split by tasks in the BioCreative PPI challenges. **AC**: article classification; **PN**: protein normalization; **ID**: interaction detection; **ME**: method extraction.

any article is registered with an empty result set, no matter the error reason. If the article is successfully received by an annotation server during the test phase, the server has to respond with results within ten minutes and the article is removed from the queue. If no response is received, the BCMS registers the result as an empty annotation set for that server and team during the test phase. Finally, responses are probed for validity (syntax and existing IDs). Malformed responses are logged and each individual annotation that could not be validated is dropped from the result set for that article.

4.2.4 Participation

In total, during all three BioCreative PPI challenges, 52 research groups participated in the four tasks presented in this work (article classification, protein normalization, interaction detection, and method extraction). The exact distribution of teams per task is shown in Table 4.5. However, some of the 52 teams will have had similar participants across challenges.

The participant teams that were selected for publications are shown in Table 4.6. Concerning these participants, the following teams are represented by similar researchers across different challenges:

- 4 (II) ⇔ 32 (II.5)
- 6 (II) ⇔ 90 (III)
- 11 (II) ⇔ 9 (II.5) ⇔ 81 (III)
- 40 (II) ⇔ 37 (II.5) ⇔ 65 (III)
- 42 (II) ⇔ 42 (II.5)
- 31 (II.5) ⇔ 89 (III)

Team	BC	AC	PN	ID	ME	Reference
4	II	Y	Y	Y	N	(Baumgartner et al. 2008)
6	II	Y	Y	Y	N	(Alex, Grover, Haddow, Kabadjov, Klein, Matthews, Tobin & Wang 2008)
11	II	Y	Y	Y	N	(Abi-Haidar et al. 2008)
†28	II	Y	Y	Y	N	(Huang et al. 2008)
40	II	N	Y	Y	Y	(Rinaldi et al. 2008)
42	II	N	Y	Y	N	(Hakenberg, Plake, Royer, Strobelt, Leser & Schroeder 2008)
9	II.5	Y	N	N	n/a	(Kolchinsky et al. 2010)
†10	II.5	N	Y	N	n/a	(Dai, Lai & Tsai 2010)
†16	II.5	Y	N	N	n/a	(Lan & Su 2010)
†18	II.5	N	Y	Y	n/a	(Chen et al. 2010)
†22	II.5	N	Y	Y	n/a	(Saetre et al. 2010)
31	II.5	Y	Y	Y	n/a	(Cao et al. 2010)
32	II.5	N	Y	Y	n/a	(Verspoor et al. 2010)
37	II.5	N	Y	Y	n/a	(Rinaldi et al. 2010)
42	II.5	N	Y	Y	n/a	(Hakenberg et al. 2010)
65	III	Y	n/a	n/a	Y	(Schneider et al. 2011)
†73	III	Y	n/a	n/a	N	(Kim & Wilbur 2011)
81	III	Y	n/a	n/a	Y	(Lourenco et al. 2011)
89	III	Y	n/a	n/a	Y	(Agarwal et al. 2011)
90	III	Y	n/a	n/a	Y	(Wang et al. 2011)

Table 4.6: Task participation by teams during the BioCreative (BC) challenges that published the methods and conclusions of their systems after the challenge. († denotes teams that only participated in a single challenge; see text for the other team correspondences between challenges.) **AC**: article classification; **PN**: protein normalization; **ID**: interaction detection; **ME**: method extraction; **Ref**: reference to team paper; **Y**: yes (participated); **N**: no (did not participate); **n/a**: not applicable (task not part of challenge).

Team	Run	BC	spec.	sens.	acc.	MCC	AUC
4	3	II	68%	79%	74%	0.48	70%
6	1	II	65%	86%	75%	0.52	**85%**
11	1	II	52%	**87%**	69%	0.41	79%
28	1	II	**73%**	81%	**77%**	**0.54**	84%
9	S28	II.5	**98%**	41%	**92%**	0.51	67%
16	4	II.5	94%	52%	90%	0.47	53%
20†	S32	II.5	96%	**59%**	**92%**	**0.56**	**68%**
31	2	II.5	96%	54%	91%	0.53	49%
65	2	III	93%	59%	88%	0.53	64%
73	4	III	**94%**	58%	**89%**	**0.55**	**68%**
81	S10	III	*100%	*6%	*85%	*0.18	*49%
89	2	III	82%	**77%**	81%	0.47	62%
90	3	III	94%	57%	88%	0.53	65%

Table 4.7: **Article classification** results of the best AUC PR submissions by teams that have published system papers (Table 4.6; except team 20, marked with †). Best scores per challenge in bold (except team 20). Note that team 81 (results marked with *) later reported their system setup had several errors, possibly explaining their huge performance loss compared to earlier submissions (as team 11 in BC II and team 9, BC II.5). AUC: AUC PR.

4.3 BioCreative PPI Evaluation Results

The results submitted by participants of the BioCreative PPI challenges were evaluated on each task using the corresponding metrics described in Section 3.2.5 (Materials and Methods). The settings for the challenges were described in Section 4.2.

For the remainder of this work, runs from online submissions are prefixed with an "S", followed by a unique server ID, while offline submissions are simply numbered from 1 to 3 (BC II) or 5 (BC II.5 and III); Therefore, a submission ID "3" would refer to the third offline submission of a team, whereas the ID "S34" would refer to an online submission via a team's server with the unique ID "34". The full, tabulated results are presented in Appendix A.1, and document-centric, macro-averaged results can be found in the respective overview paper for each challenge. In this work, we evaluated the global (i.e., not document-centric) results using macro-averaging[13] and present each team's best run/submission.

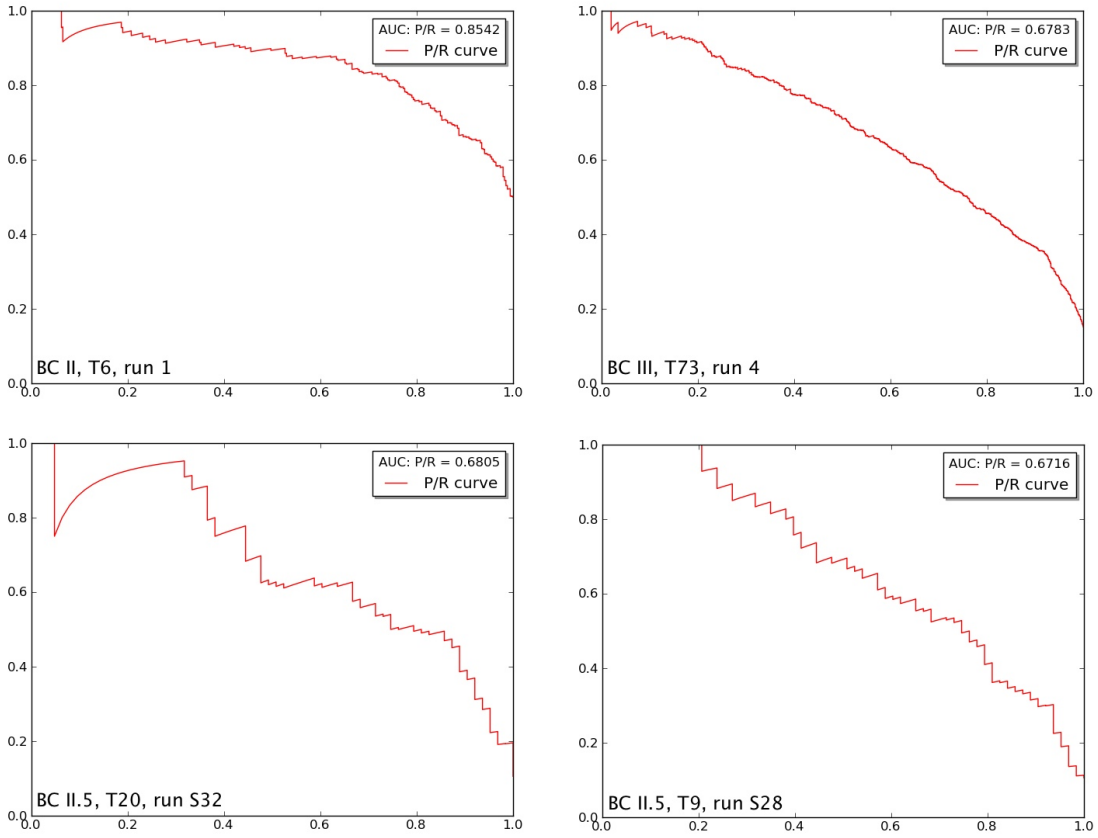

Figure 4.4: Precision/recall curves calculated from the ranked results of the best **article classification** ranking systems for each BioCreative (BC) challenge and team (T). As team 20 (BC II.5) did not publish their system, the second best result (team 9, S28), is shown, too.

4.3.1 Article Classification

The final evaluation scores were calculated for all three challenges on a total of 134 result sets submitted by 34 teams. The detailed scores per team and submission and a specificity vs. sensitivity plot can be found in Appendix A.1.1.

The best overall article relevance ranking score for BioCreative II was 85.4% AUC PR (team 6, run 1); the best classification run was submitted by team 28, run 1 (MCC = 0.544). The best BioCreative II.5 ranking approach was submitted by team 20 during BioCreative II.5 (S32, 68.1% AUC PR). Despite the result, team 20 (K. Ambert and A. Cohen) was not willing to publish its system; However, the second best result with 67.2% AUC PR by team 9 (S28) is published, and this team had the best overall MCC score (0.583, with S29) across all challenges. The best scores measured in BioCreative III were 67.8% AUC PR (team 73, run 4) and a MCC of 0.553 (team 73, run 2). PR curve plots for the best ranking results are shown in Figure 4.4. The best runs of the tasks' participants listed in Table 4.6 are shown in Table 4.7.

4.3.2 Protein Normalization

A total of 94 normalization results sets were submitted by 27 teams during BioCreative II (46 runs) and II.5 (48 runs). For the baseline evaluation presented in Figure 4.5, the entire UniProt annotations were used from both challenges. Thus, the results are directly comparable. However, during BioCreative II, most participants did not (correctly) rank their results and were not asked by the organizers to report normalized confidence values for each result, so that a ranking evaluation (AP and FAP-score) cannot be made for the results of the BC II normalization and interaction tasks. Only BioCreative II.5 teams are shown in the plot comparing AP and F-score in Figure 4.6, and only for the BC II.5 submissions the AP and FAP scores are given for the detailed results presented in Appendix A.1.2.

The main evaluation target for the BioCreative II challenge was the F-measure. For BioCreative II.5, teams were instructed that their ranking performance would be evaluated via AP as primary evaluation function, while F-measure was applied as a secondary evaluation target. Some teams, in particular, 10, 22, and 51, chose to maximize AP performance that lead to huge result sets with large recall but minimal precision reported by those teams, as can be seen in Figure 4.5. The best overall F-measure was achieved by team 40, run 2, with an F-measure of 29% at 29% recall and 33% precision during

[13] except for the article classification task, where this distinction is non-relevant

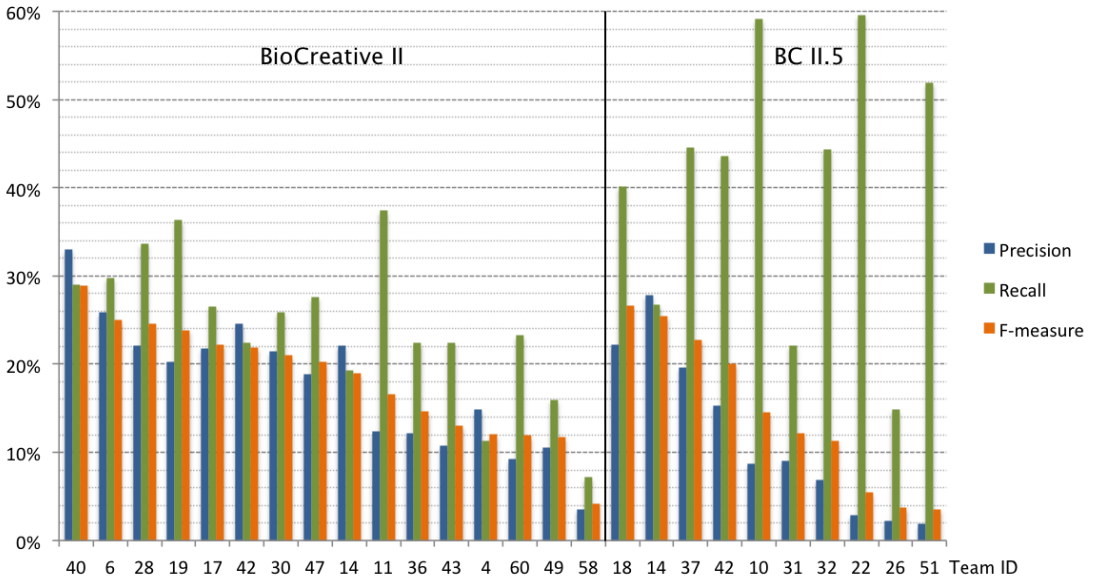

Figure 4.5: Macro-averaged **protein normalization** results of the best submission (in terms of F_1) per team, ordered by (macro-averaged) F-scores. Team-Run, BioCreative II: 40-2, 6-3, 28-1, 19-2, 17-3, 42-3, 30-3, 47-2, 14-1, 11-3, 36-1, 43-1, 4-1, 60-1, 49-2, 58-3. Team-Run, BioCreative II.5: 18-5, 14-1, 37-3, 42-S02, 10-S09, 31-S18, 32-1, 22-3, 26-S16, 51-3.

BioCreative II. The best BioCreative II.5 run (team 18, run 5) achieved 27% F_1 (40% recall, 22% precision). The best AP score (BC II.5 only) of 26% was achieved by team 10, run S09, followed by team 22, run 3, with 23%. The AP curves of these two teams are plotted in Figure 4.7.

FAP-Score Cutoff

To establish the maximum F-score the high-recall runs could hypothetically achieve, a new metric is introduced in this work, the FAP-score[14]. For the three teams that optimized their systems against AP, the confidence cutoff that maximizes FAP was established on the *training* set data. We searched for the confidence score that produces the highest FAP-score on the training data using 0.01 score increments.

For team 10, run S09, this confidence score is 0.04 on the training set data, and for team 22 it was 0.71. Team 51 reported the same confidence value for all runs (1.0), so the best cutoff rank was established (rank 11). The performance values of these adjusted result sets are shown in Figure 4.6 as orange dots. Two of these adjusted runs surpass the best official F-measure result (29%), with F-measures of 39% (team 10, at 46% recall,

[14]see Section 3.2.5, Materials and Methods, as the harmonic mean of F-measure and AP

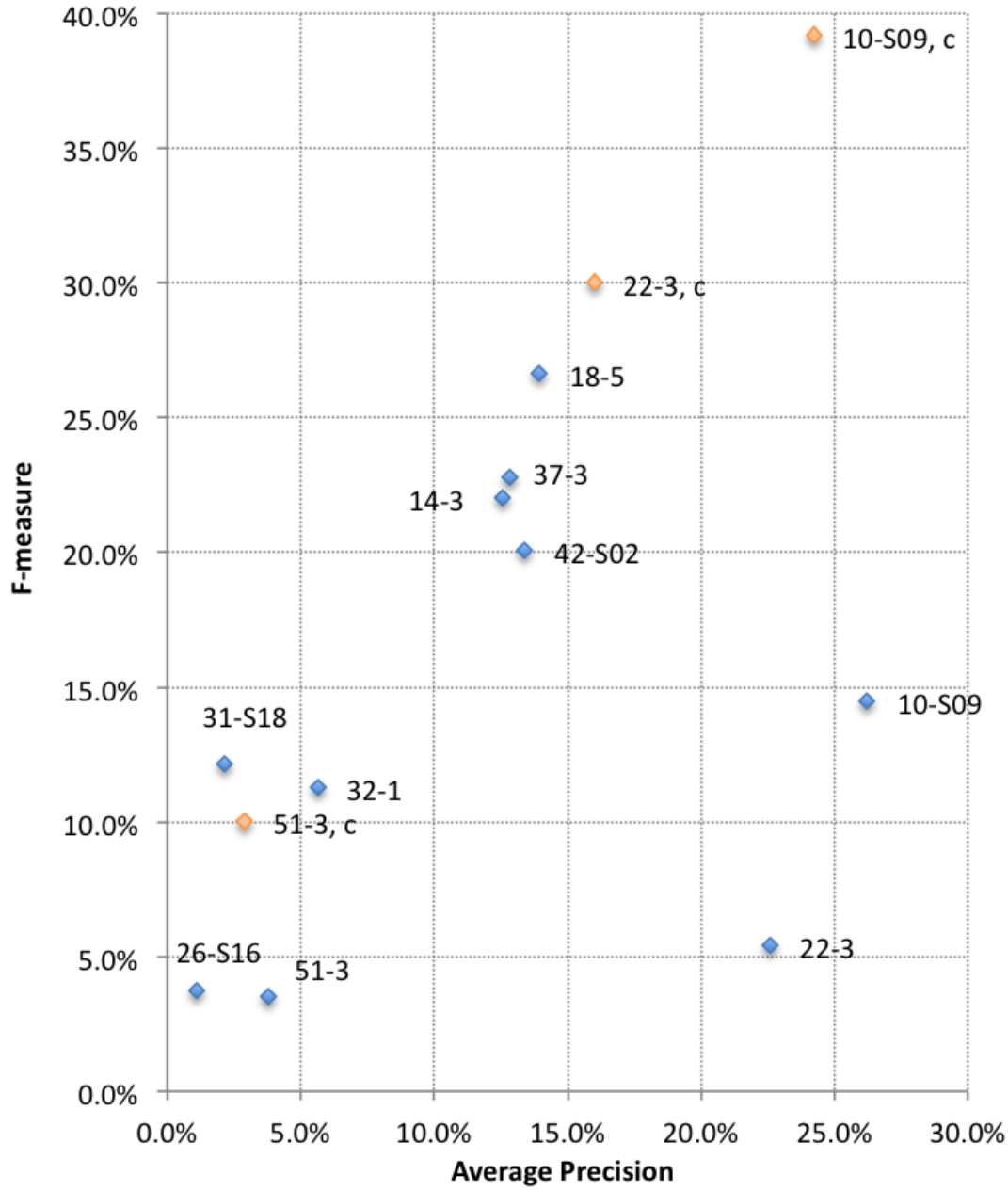

Figure 4.6: Macro-averaged AP vs. F_1 comparison of the best AP **protein normalization** results per team and run (labeled team ID - dash - run ID) in the BioCreative II.5 challenge (only). In addition, the orange dots indicate possible results created by establishing a confidence/rank cutoff for the high-recall runs at their maximum FAP-scores on the *training* set that maximizes their FAP-score.

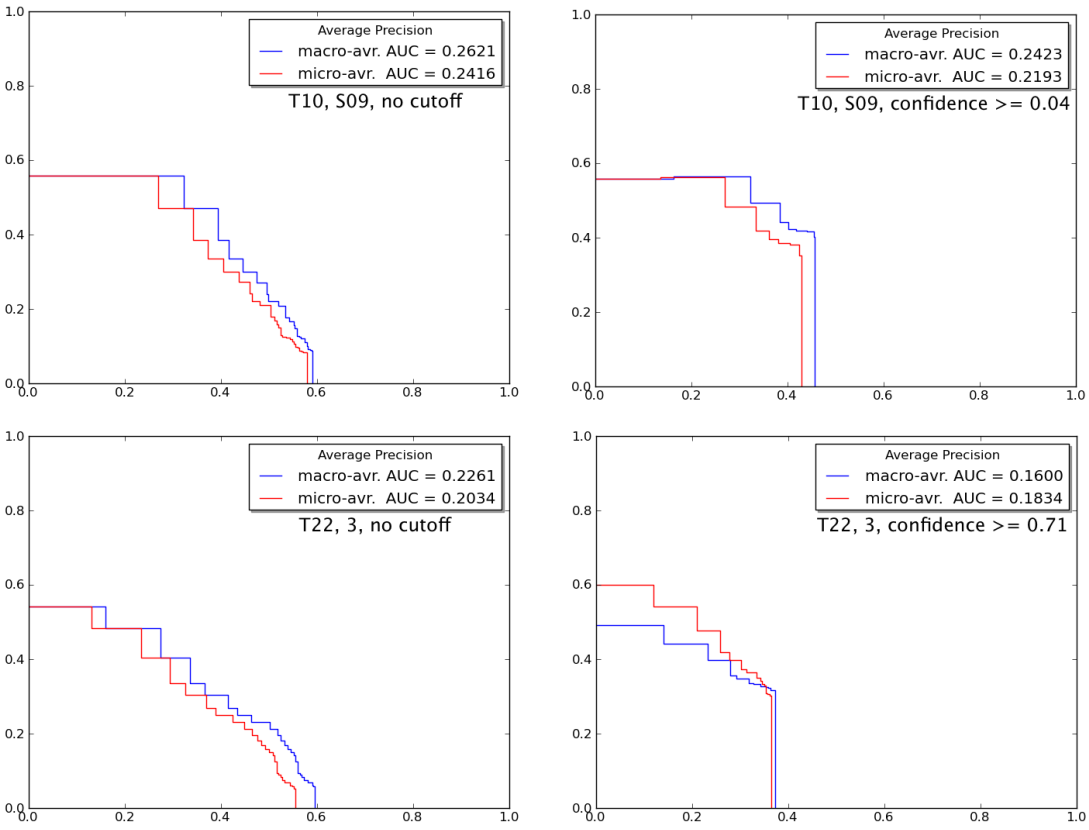

Figure 4.7: AP curve plots of the two best ranked **protein normalization** results of teams 10 (top, left) and 22 (bottom, left) during BioCreative II.5. To the right, the AP curves of these two results are shown after applying the FAP-based training set cutoffs.

39% precision) and 30% (team 22, 37% recall, 31% precision). To illustrate the impact, on run 10-S09, this adjustment reduces the annotations reported by the system from 1784 to 348 (i.e., to one fifth). Team 10 also achieved the highest FAP score of 19% with their (original) run, closely followed by team 18 (FAP = 18%). However, teams 42, 37, and 14 (all 16% FAP) all have a higher FAP-scores than team 22 (FAP = 9%). The hypothetical cutoff result sets would have had FAP scores of 30% (10-S09) and 21% (22-3). The effect of taking these optimal cutoffs on the AP curves of teams 10 and 22 are displayed in Figure 4.7 (right side).

SwissProt vs. UniProt

In addition, the impact of evaluating against SwissProt only articles and of only regarding SwissProt results was measured. For BioCreative II, participants (with the exception of team 6) only normalized against SwissProt, while for BioCreative II.5, all participants normalized against the entire UniProt space. If evaluating the results on the 242 articles in the BioCreative II pair corpus that have only SwissProt accessions, the average F-measure of all best runs submitted by the teams increases by 3.3pp (std. dev. = ± 1.5), for the 46 SwissProt only articles in BioCreative II.5, the average F-score increase of the BioCreative II.5 systems is 2.1pp (std. dev. = ± 1.7). Second, the best submissions of all participants in BioCreative II.5 were evaluated after removing any TrEMBL accession in their results. The average F-measure when only using SwissProt accessions reported by the teams decreased by 1.6pp (std. dev. = ± 1.6). In the case of the best run (18-5), the decrease is even 5.4 pp (a 20% loss). For the adjusted runs (that were not included in the 1.6pp F-measure decrease average) 10-S09 and 22-3, the losses are 3.0 (a 9% loss) and 4.4pp (a 15% loss), respectively. The increase in precision from removing the TrEMBL accessions does not equilibrate the decreased recall, and that is especially true for good normalization systems.

4.3.3 Interaction Detection

For the interaction detection tasks of BioCreative II and II.5, a total of 84 runs were submitted by a total of 24 teams (16 teams in BC II, 8 in BC II.5). Just as for the normalization task, no ranking data for the BioCreative II results was used and no AP or FAP scores are taken. The baseline evaluation results for the best F-measure submissions are shown as histogram in Figure 4.8. The best AP results from BioCreative II.5 are plotted against F-measure in Figure 4.9. The detailed results are listed in Appendix

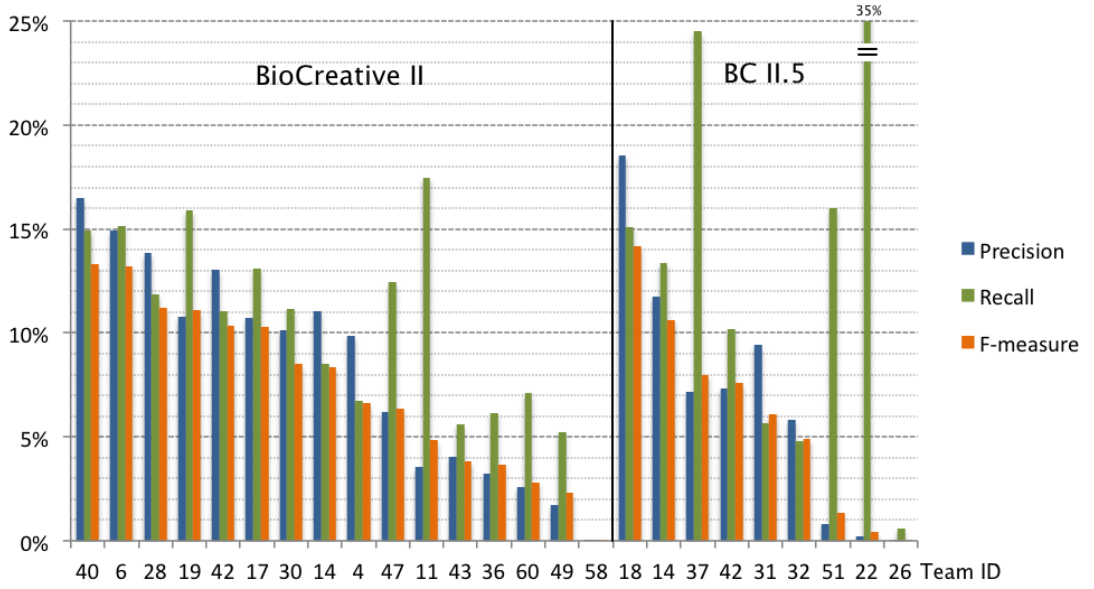

Figure 4.8: Macro-averaged **interaction detection** results of the best submission (in terms of F_1) per team, ordered by (macro-averaged) F-scores. Team-Run, BC II: 40-2, 6-1, 28-3, 19-3, 42-2, 17-3, 30-3, 14-1, 4-1, 47-2, 11-3, 43-3, 36-3, 60-1, 49-3, 58-3. Team-Run, BC II.5: 18-5, 14-1, 37-4, 42-S01, 31-5, 32-S07, 51-3, 22-1, 26-S16.

A.1.3.

For BC II.5, the best overall F-measure was achieved by team 18, run 5, with an F-measure of 14% at 15% recall and 18% precision during BioCreative II.5. The best BioCreative II run (team 40, run 2) achieved 13% F_1 (15% recall, 16% precision). The best AP score (BC II.5 only) of 4.6% was achieved by team 22, run 1, followed by team 14, run 3, with 3.6%. Some BC II.5 teams, in particular, 22, 37, and 51, optimized AP performance that lead to result sets with large recalls reported by those team, as can be seen in Figure 4.5.

FAP-Score Ranking

Team 18 has the best FAP-score, with 5.1%, together with 14-3, while team 22, run 1, only achieved a FAP-score of 0.75%. As a matter of fact, the FAP-score of team 22 only puts them ahead of teams 26, 32, and 51. Even a FAP-score based confidence cutoff on the training set would not have improved their FAP-score to the levels of the three top teams[15] 18, 14, and 37. 37-4 has a FAP-score of 4.1% (3rd after teams 18 and 14), and

[15] with an optimal training set-based confidence cutoff at 0.99, their test set FAP-score increases to 3.2%

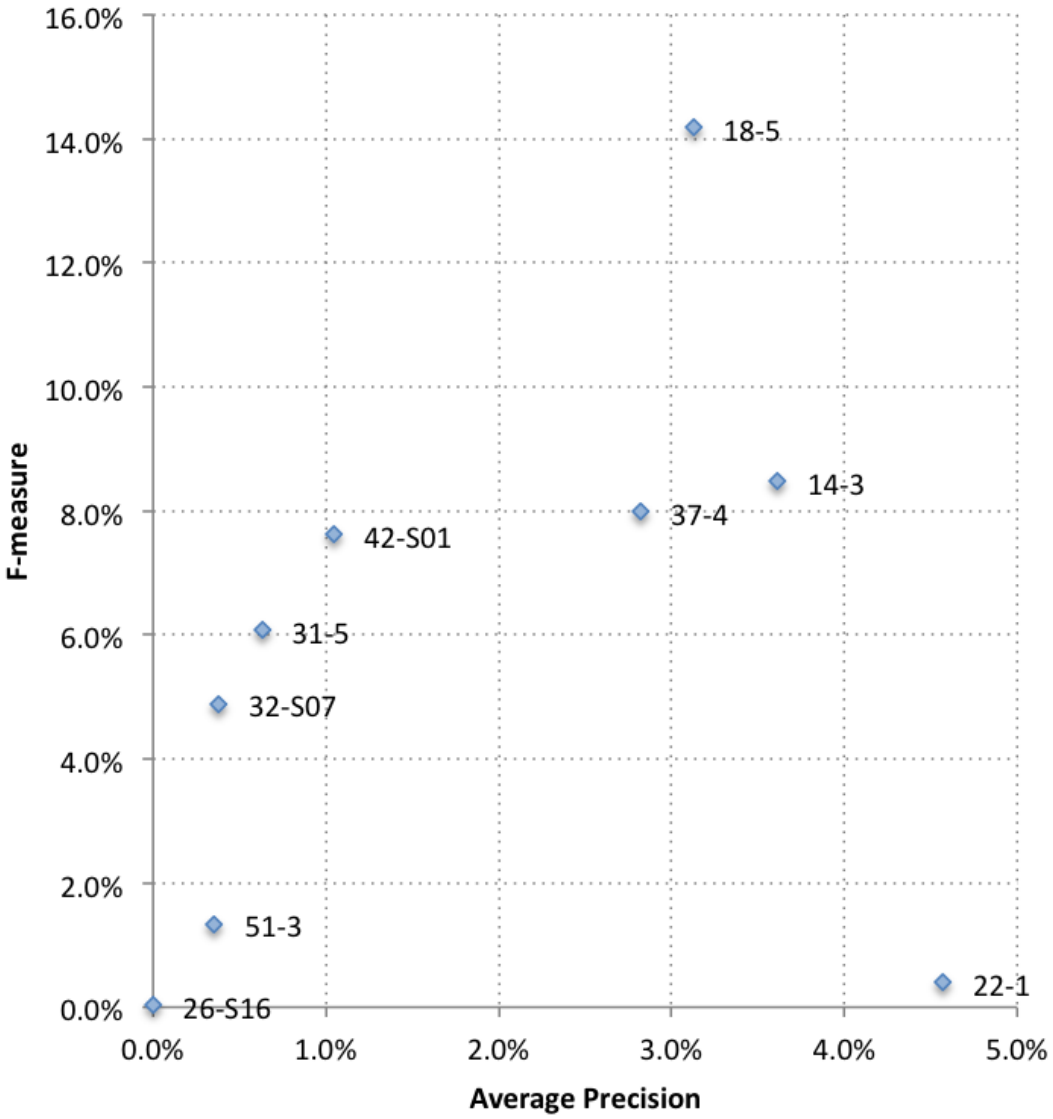

Figure 4.9: Macro-averaged AP vs. F_1 comparison of the best AP **interaction detection** results per team and run (labeled team ID - dash - run ID) in the BioCreative II.5 challenge (only).

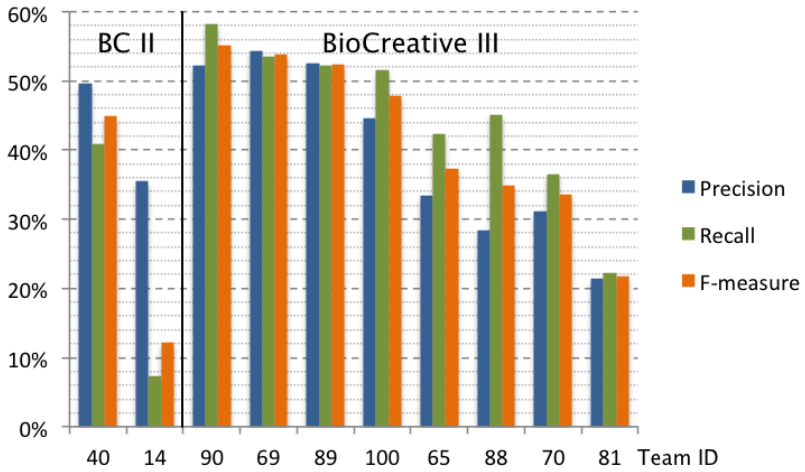

Figure 4.10: *Micro*-averaged **method extraction** results of the best submission (in terms of F_1) per team, ordered by (macro-averaged) F-scores. Teams 40 and 14 participated in BioCreative II. All other teams and results are from BioCreative III. Team-Run: 40-3, 14-1, 90-3, 69-2, 89-5, 100-1, 65-4, 88-1, 70-4, 81-5.

the FAP of 42-S01 was measured as 1.3%.

4.3.4 Method Extraction

In total, ten teams submitted a total of 48 result sets - six sets from two teams in BioCreative II, and 42 from eight teams in III. Apart from the fact that the settings for BioCreative II were more difficult[16], the scores otherwise are directly comparable. The baseline evaluation results for the best F-measure submissions by team are shown in Figure 4.10 as histogram. The best AP results from BioCreative III are plotted against F-measure in Figure 4.11. The three best team's AP curve plots plus team 65 are shown in Figure 4.12.

The best F-score (55%) and AP (35%) was achieved by team 90, run 3, measured in BioCreative III (at 58% recall and 52% precision). This run is closest to the "sweet spot" in the AP vs. F-measure plot, slightly ahead of the submissions by team 69 and well ahead of team 89. The FAP scores of these teams are 43% (team 90), 42% (team 69) and 38% (team 89). In BioCreative II, where only two teams participated in this task, the best F-score was 45% with 41% recall and 50% precision (team 40, run 3). Team 65 in BioCreative III is largely the same as team 40 in BioCreative II, i.e., they had a lower performance in the second edition of this task.

[16] because of the lower coverage of test set classes in the training set; see Section 4.1.3

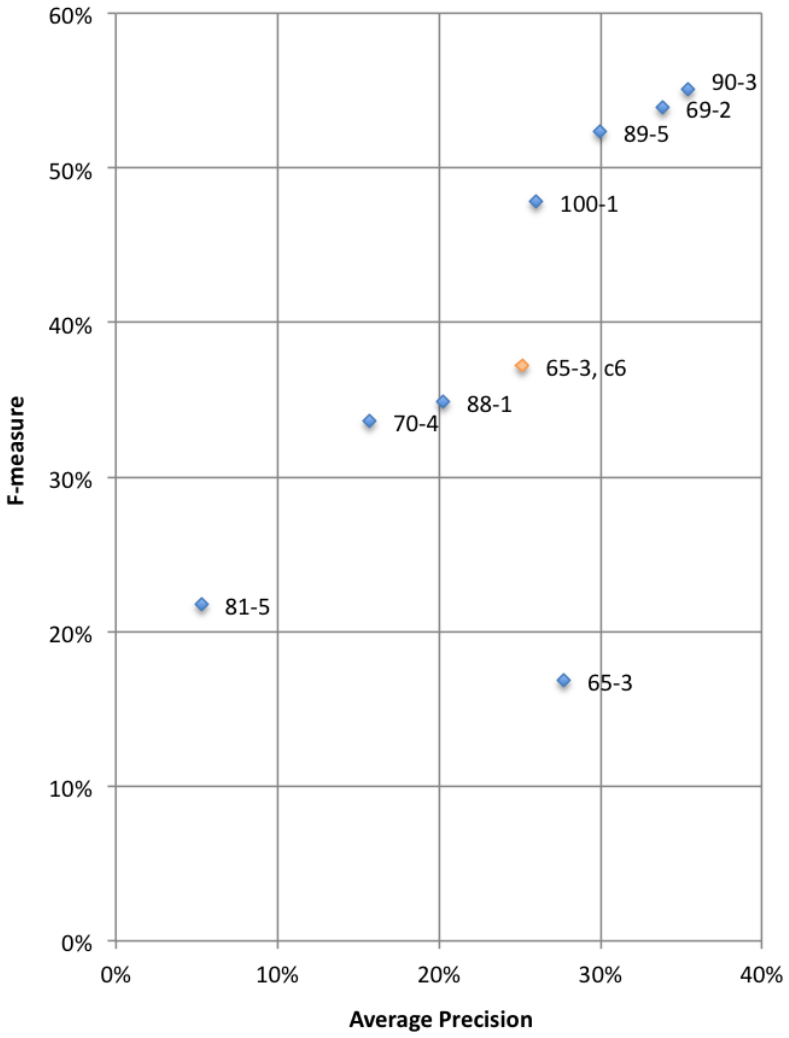

Figure 4.11: *Micro*-averaged AP vs. F_1 comparison of the best AP **method extraction** results per team and run (labeled team ID - dash - run ID) in the BioCreative III challenge (only). In addition, the orange dot indicates an *artificial* result for the run of team 65. The result was created by cutting of their run at rank 6.

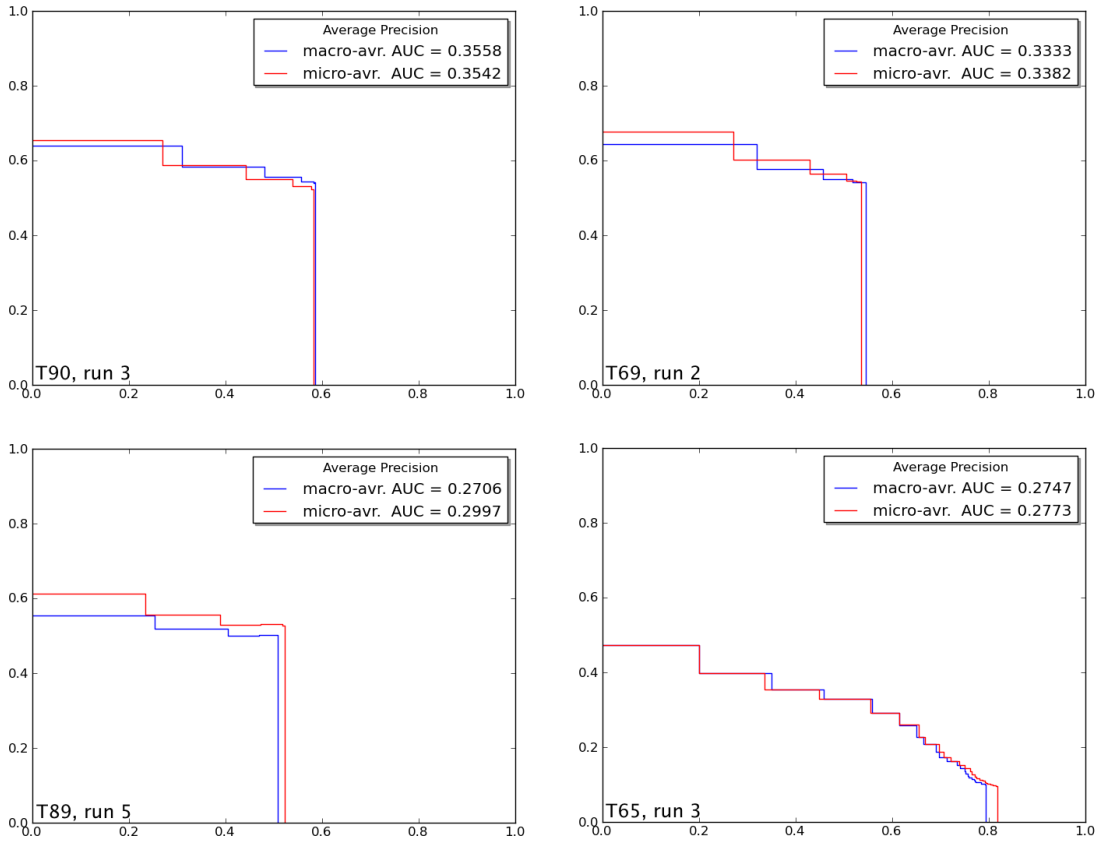

Figure 4.12: AP curve plots of the three best ranked **method extraction** results of teams 90 (top, left), 69 (top, right) and 89 (bottom, left) as well as the submission of team 65 (bottom, right) with an artificial rank cutoff during BioCreative III.

FAP-Score Cutoff (Team 65)

The FAP-score was used to find the optimal cutoff for team 65, that optimized its classifier for AP (see Figure 4.12, bottom right) with run 3. The cutoff that maximized the micro-averaged FAP was at rank 6 in the test set. No training set runs were submitted during the BC III challenge, therefore no comparison to a training set cutoff can be made, and this result is entirely artificial. The performance of this artificial result set is shown in Figure 4.11 as orange dot. It has the same F-score as their best submission by F-measure (run 4, see histogram in Figure 4.10) in BioCreative II.5, but increases the FAP of run 3 from 21% to 30% (Run 4 of team 65 had a FAP of 19%.) However, their AP decreases from 28 to 25%, putting them behind team 100 (AP = 26%) in the team order on ranking performance evaluation, as can be seen from the position of the orange dot in Figure 4.12. This is the only case where a FAP-based adjustment changed the team order based on AP when this procedure was applied. Although in this case the adjustment is artificial, it is an example that shows that AP can be substantially "boosted", particularly when the number of possible classes is small (here, the 115 possible PSI-MI concepts).

4.3.5 Document-centric Organism Filtering Evaluation

All protein normalization and the BioCreative II.5 interaction detection results were further analyzed using document-centric evaluation after applying the organism filtering procedure as described in the Materials and Methods (see Section 3.3.3). To measure the impact of organism mapping errors or choosing wrong focus organisms for the normalization step, the impact of removing proteins for these irrelevant organisms is measured on each team's best run. The official results reported in the challenge overview papers are measured using this document-centric evaluation, which increases recall. The organism filtering, on the other hand, increases precision. Their combined application evaluates the maximum performance of a system for all cases where the system did produce results on the article after a (human) user has selected the relevant organisms.

The document-centric *protein normalization* runs of each team after applying the organism filter are shown in Figure 4.13. Under this scenario, the best normalization approach was submitted by team 6 (run 3) during BioCreative II with an F-score of 48%[17]. However, this result only produced annotations for 199/346 documents (58%).

If going beyond the best baseline runs of each team[18], the FAP-score adjusted run of

[17] prec. = 53%, rec. = 52%, all macro-averaged

[18] i.e., taking all runs into account, not just the best baseline submission of each team

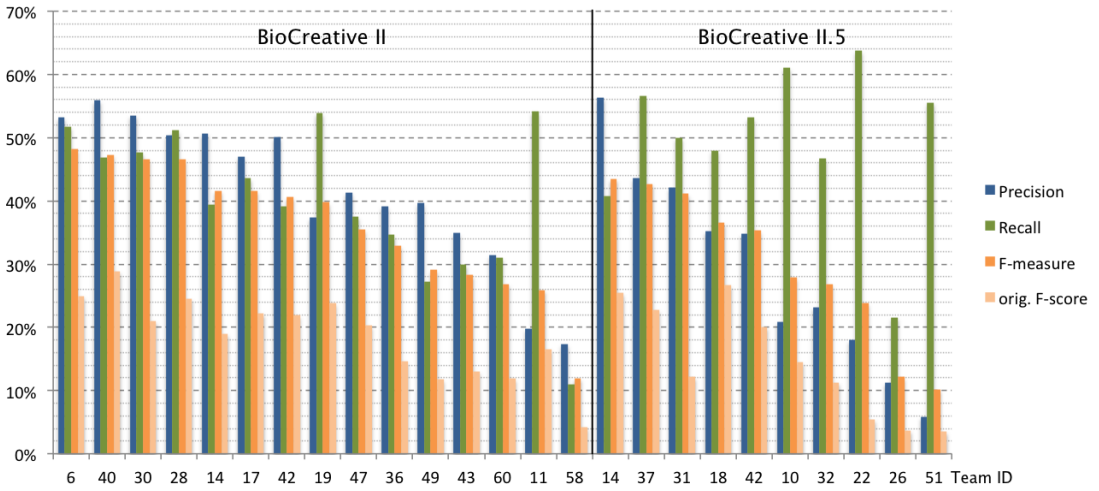

Figure 4.13: Macro-averaged, *document-centric* **protein normalization** results after *organsim filtering* of the best submission (in terms of F_1) per team, ordered by (macro-averaged) F-scores for the same runs as shown in Figure 4.5.

team 10 (S09) covers 59/61 documents (97%), while it has the same F-measure of 48%[19]. The AP curve plot of this run is shown in Figure 5.4 (in the Discussion). Furthermore, the best organism-filtered run in all BioCreative challenges was submitted by team 42, BC II.5, with S01 (Their best baseline was 42-S02). While this run only covers 19/61 documents (31%), the F-measure of this run was 56%, with a precision of 68% and a recall of 53%.

The document-centric *interaction detection* runs from BioCreative II.5 after applying the organism filter were measured, too. Under the document-centric, organism-filtered scenario, the best BC II.5 interaction detection approach was submitted by S01 of team 42 (BC II.5, F-score: 35%) across both BC challenges. The second and third best runs were 18-5 and 32-3, respectively (see Figure 4.15 with their scores).

4.3.6 Author-Curator-System Comparison

In addition to assessing the performance of automated PPI extraction methodologies, BioCreative II.5 was designed to compare the systems with the performance of curators and authors on the identification of the (binary) interaction pairs (interaction detection task) and the underlying identification of the interacting proteins described in the article (protein normalization task). The curator and author results were generated in the context of the FEBS Letters experiment on SDAs (see Introduction, Section 1.3.5). For

[19] prec. = 56% prec., rec. = 47%

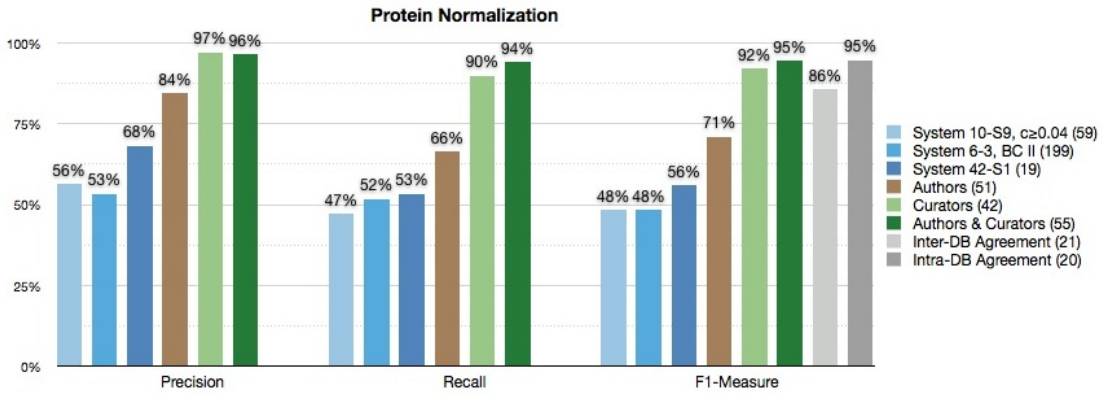

Figure 4.14: BioCreative II (team 6) and II.5 (team 10 and 42) **protein normalization** results from the three best-performing automated systems after document-centric organism filtering (blue), the authors (brown), and the curators (green, light: curator alone; green, dark: using author annotations), as well as the IAA (grey, light: inter-DB; grey, dark: intra-DB). Legend numbers in parenthesis: articles annotated.

the comparison, the organism-filtered system results were used, as especially the authors, but possibly curators, too, would have no trouble selecting the correct organisms. These direct comparisons between systems, authors, and curators are shown in Figures 4.14 (normalizations) and 4.15 (interactions). The performance of each source (systems, authors, curators) was evaluated in terms of precision, recall and balanced F-measure (the harmonic mean between precision and recall).

The DB curators also generated annotations based on author data (authors & curators), i.e., instead of creating annotations from scratch, they based their annotations on the author-supplied data (dark green curator results). To measure the Inter-Annotator Agreement (IAA), three curators, two from the same PPI database (intra-DB) and another from a distinct database (inter-DB), created annotations on the same articles. The annotation data from these three curators on their commonly annotated articles was used to measure the IAA of the curators' annotations and is depicted by the grey bars in Figures 4.14 and 4.15 (light grey: inter-DB, dark grey: intra-DB agreement).

The best protein normalization system's top ranked results from the *training*[20] data (and without any organism filtering) per article were used to calculate the annotation overlaps shown in Figure 4.16. In total, there are 33 articles the three sources - the

[20]The system that was used in this analysis - team 10, S09 - had the best scoring AP and FAP on the test set, but performs 4pp (21%) worse on the test set (macro-averaged F-score = 15%) compared the training set (F_1-measure = 19%) . As the SDA annotations were publicly available at the time of the BioCreative challenge, the automated systems could not be tested on the articles annotated by the authors and were evaluated on an independent test set, as described in Materials and Methods, Section 3.1.2.

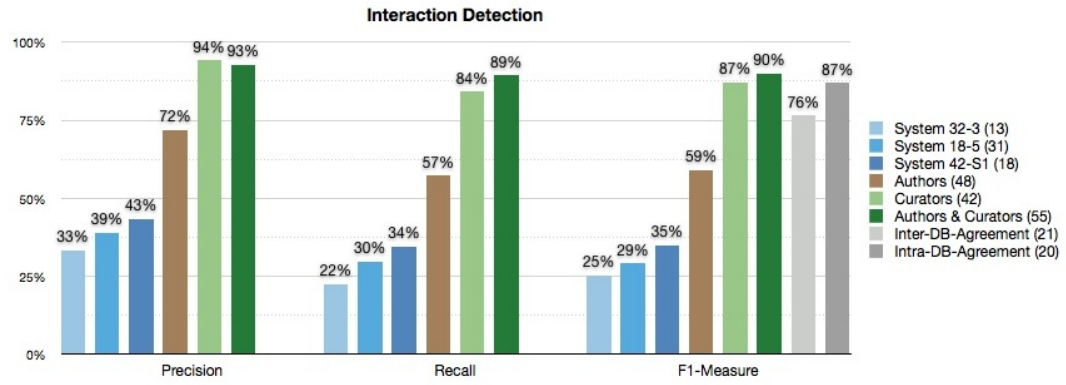

Figure 4.15: BioCreative II.5 organism-filtered **interaction detection** results from the three automated systems selected by best F-score (blue, after organism filtering) compared to the authors (brown), the curators (green, light: curator alone; green, dark: using author annotations), and the IAA (grey, light: inter-DB; grey, dark: intra-DB). Legend numbers in parenthesis: articles annotated.

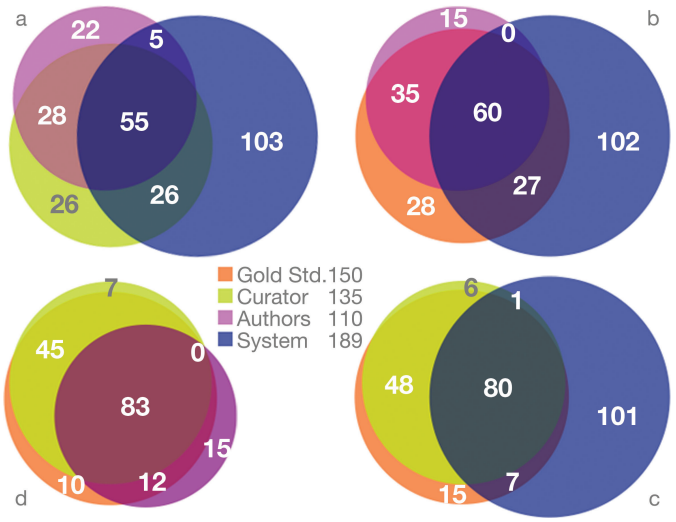

Figure 4.16: Overlap of UniProt accession annotations by authors, curators and an automated system (team 10, S09), shown as Venn diagrams. The articles are the 33 (training set) articles that all three sources annotated in common. (Clockwise) Top left, a: Comparison of the result sets from all three sources of annotations. All others visualize the set overlaps relative to the gold standard: Top right, b: Authors and text-mining systems. Bottom right, c: Curators and text-mining systems. Bottom left, d: Curators and authors.

system, authors, and the curators - had annotated in common. The automated system, for each article, produces a relatively long list of proteins ranked according to relevance. For the comparison with the human annotators only the system's top six proteins for each article are considered[21] (see Figure 4.7).

4.3.7 Author Questionnaire Responses

From March to December 2008, 76 authors were invited to take part in the FEBS Letters SDA experiment. Three quarters (57) accepted and the articles were published with an appended Structured Digital Abstract (SDA) reporting in a human readable format (1) the identifiers of the interacting proteins, (2) the interaction type (physical interaction, co-localization, enzymatic reaction, etc.), and (3) the method used to support the interaction. The quarter of authors that did not accept explained that they were worried to delay the publication of their articles or preferred to avoid complying with this extra request because of other commitments. After publication, the participant authors that had returned curation-relevant annotations were asked to fill a simple questionnaire with ten questions, but only 22 of those responded to the questionnaire.

Summarizing the survey results from these 22 authors, the majority of the participants demonstrated interest in the experiment although seven complained that insufficient information was given to efficiently complete the task. 17 were very supportive but two authors explicitly said that they would be discouraged from publishing in FEBS Letters if the procedure were made compulsory. Seven authors complained about insufficient guidelines for creating the annotations. Authors only received an Excel form that contained the guidelines for creating the annotations. The authors declared that this extra responsibility took, on average, 68 minutes of their time (std. dev. $= \pm$ 72 min) and that the most demanding task was the retrieval of the protein identifiers from UniProt. In conclusion, slightly more than one quarter of the authors objected to the extra task of creating structured data for their publications or showed reluctance to performing this task again in the future.

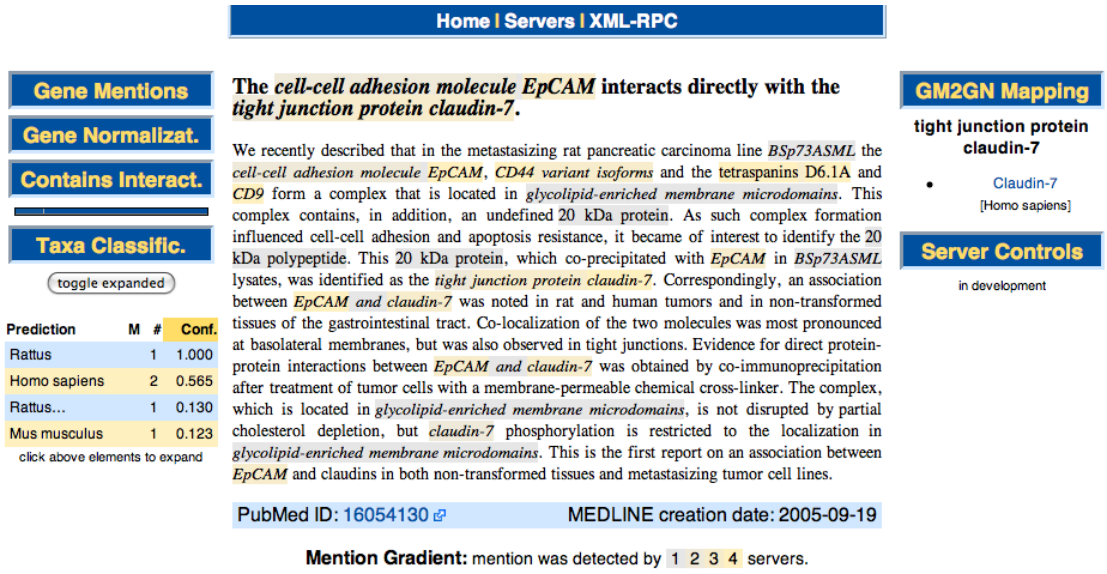

Figure 4.17: Screen-shot of the BCMS displaying the annotation results for PubMed ID 16054130. **Left column:** The annotation result browser, with the taxa classification results shown expanded and sorted by confidence. In addition, the bar below "Contains Interact." indicates by color (ranging from yellow to blue) if an abstract is likely to contain PPIs, here a clear vote (i.e., blue) by the annotation servers that the publication describes PPIs. **Central area:** The abstract itself, but with the protein mention annotations highlighted as heat-map from gray (only one server reported the annotation) to yellow (all four servers report the text span as a gene/protein mention); As can be seen, the relevant names (EpCAM, claudin-7, CD-44, D6.1A, and CD9) are all annotated by all servers. **Right column:** Mention-normalization mapping view; After clicking on a highlighted mention in the text, and if a name mapping between the mention and a UniProt normalization reported by any of the servers exists, the relevant link to the UniProt entry is shown.

4.4 BioCreative Meta-Server

4.4.1 Silver Standard Annotations

The first prototype of the BioCreative Meta-Server[22] (BCMS) created after the BioCreative II challenge collected results for 22,730 PubMed abstracts from annotation servers provided by thirteen research groups. These annotation servers provide the BCMS with one or more of four principal annotation types for each PubMed abstract:

1. Gene/protein mentions in the abstract as offset values[23] in the text
2. Gene/protein IDs[24] mapped to the abstract
3. Taxonomic IDs[25] mapped to the abstract ("focus organism")
4. PPI - a Boolean value indicating if the abstract contains PPI information or not

Each of these annotation types is reported together with a normalized value to indicate the annotation system's confidence in the classification. The abstracts were annotated with over 1.6 million[26] protein/gene mentions, 237,600 protein/gene identifiers (for UniProt and Entrez Gene), 1,376 species identifiers (NCBI Taxonomy), and are classified on average by 4 different systems as PPI-relevant or not. All annotations are provided with a confidence score in the range $(0, 1]$.

4.4.2 Public Web Interface

While the BCMS communicates entirely via XML-RPC web services with both the annotation servers and the public, the entire data can be inspected via browser, too. The technical details and components have been presented in Materials and Methods, Section 3.5. A screenshot of the web interface is shown in Figure 4.17. Mapped proteins and genes are hyperlinked to the corresponding public database records of those resources.

[21] If applying the optimal confidence cutoff (confidence ≥ 0.04) on the *training* data of 10-S09, the training run achieves an (macro-averaged) F-score of 51%; using a rank cutoff (rank ≤ 6), the F-score decreases to 42%, incurring an 9pp (18%) F-score loss. This mimics the 21% F-score decrease when comparing the training and test set F-scores of run 10-S09.

[22] `http://bcms.bioinfo.cnio.es/`

[23] I.e., positionial information about a sub-string. For example, in the string "abc", "a" has an offset of zero, "b" of one, and "c" of two. The substring "ab" is found at offsets 0:2.

[24] Entrez Gene IDs and UniProt accessions

[25] NCBI Taxonomy IDs

[26] 1,659,219

Upon receiving a request for a given abstract from the web interface, the BCMS immediately shows any already stored results and presents the abstract itself. Once more results are received from annotation servers, the web interface automatically updates with the latest results[27]. The BCMS then stored all received (and pruned) annotation data.

For gene/protein mentions, each letter tagged by one or more servers is highlighted with a gradient indicating the number of annotations made on that letter. For the protein and gene normalizations, all official names and synonyms are fetched for each record and regularized. Regularization converts names to lower case, and removes white-spaces and dashes. If any of these strings match a (regularized) mention tagged by the gene/protein mention servers, this mapping is presented as a possible link between the protein/gene normalization and mention. All annotation sets can be ordered by confidence values or by the number of servers that have made the corresponding annotations.

4.4.3 Reactor Pattern Implementation

Given that the BCMS at the same time needs to communicate with clients at the front-end and itself assumes a client role when communicating with annotation servers at the back-end, a critical issue was coupling the two processes. A single front-end request might trigger a back-end request to each annotation server, and the BCMS must be able to accept multiple, concurrent front-end requests at the same time.

However, the back-end annotation severs are not expected to be capable of multi-threading, i.e., most of them can only handle one request at a time. Furthermore, the BCMS has to be able to handle connection failures and time-outs, communications errors, and data validation issues. Therefore, the BCMS back-end works in parallel mode relative to each server, but each server thread is modeled as its own serial process. Put slightly differently, the BCMS acts as a "buffer" between public requests and the annotation servers.

The main back-end design is based on a persisted queue for each server that can be filled up by incoming front-end requests or requests can be added manually. The BCMS launches one back-end request to each annotation server at a time. Once a server responds with its annotations, the BCMS validates and stores the results. Then the BCMS immediately supplies the server with a new request if that server's queue is not empty. If a front-end request is waiting for that particular result, it can be sent as soon

[27]It should be noted that by now all servers have completed their annotations, so no more back-end updates are occurring, and the servers are listed as "offline".

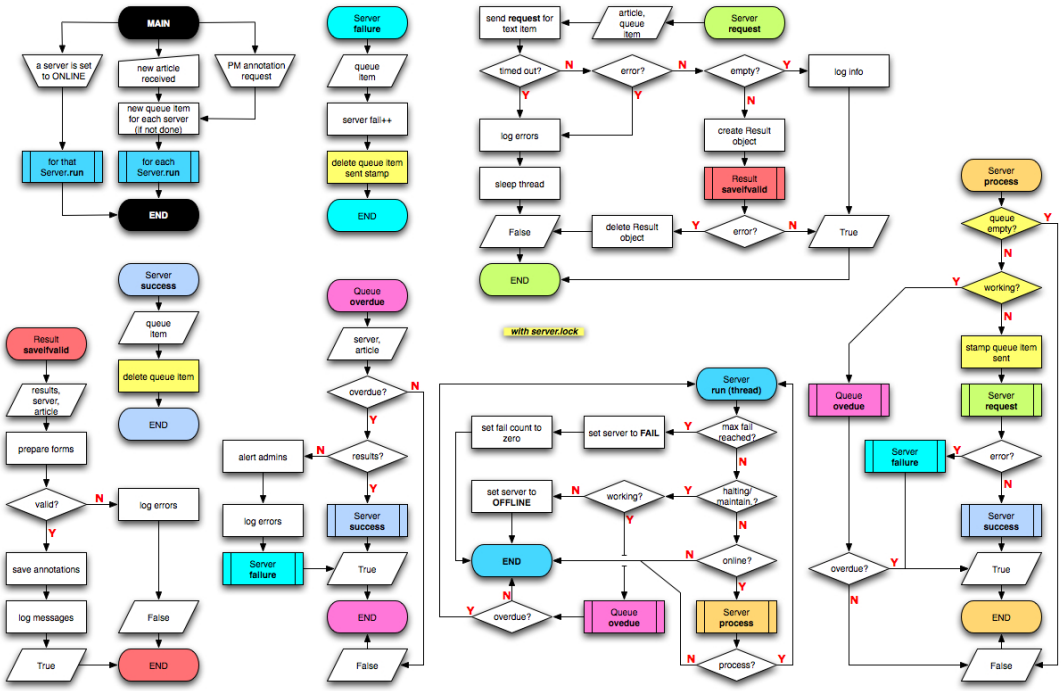

Figure 4.18: Process diagram of the main back-end process of the BCMS. The black MAIN loop implements the reactor. See Table 4.8 for a more detailed description of each process. Yellow steps highlight the placement of thread-locks.

Subroutine, Figure Color	Description
Server.run (thr.), blue	The main thread loop executed for each annotation server, resulting in (Server.process, orange) calls or time outs (Queue.overdue, purple). Depending on the state of the server (ONLINE or HALTING) and the number of errors that occurred, the next queue item might be processed (ONLINE) or the state changed accordingly (to OFFLINE or FAIL), after which the server's thread is terminated.
Server.process, orange	This subroutine ensures that either a request is sent to the server (Server.request, green), reports a failure to do so (Server.failure, turquoise), or triggers a time out (Queue.overdue, purple). If the server request process (Server.request, green) signals success, the success routine (Server.success, gray) is called, otherwise the failure method (Server.failure, turquoise) is executed.
Server.request, green	The actual request is made to the server. Only if a response was received and is non-empy, the BCMS attempts to store the results (Result.saveifvalid, red).
Result.saveifvalid, red	If a response was received, the data is analyzed for correct syntactical and semantical content and, if valid, gets saved.
Queue.overdue, purple	If a server does not respond to requests for a fixed amount of time, the overdue routine gets triggered (a "call-back"). In case the outstanding request cannot be made, this results in a failure (Server.failure, turquoise).
Server.success, gray	Final routine executed after completing successful annotation requests; It deletes the annotation item from the server's queue. Because the (ACID-compliant) DB acts as a global lock, multiple BCMS instances can be run in parallel.
Server.failure, turquoise	If any error occurred, this function augments the error count for the server and removes the article from the server's annotation queue.

Table 4.8: BCMS event loop subroutine descriptions for Figure 4.18.

as the result is validated. This ensures that even if servers have entirely different time requirements to complete their annotations, each server is used at constant load (or none at all if no requests are queued) and each front-end request receives a response as fast as possible. The details of this main back-end process loop are shown in Figure 4.18 and details of the figure are explained in Table 4.8.

Starting with the main process loop (black, MAIN, top left in Figure 4.18), each annotation server can be dynamically added to and removed from the main process (left branch). From a computer science perspective, this main loop implements a *reactor pattern*[28]. Each annotation server itself is a single threaded loop ("Server.run (thread)") branched from the main process with its proprietary annotation queue. These threads represent *event loops*.

The elegance is that while the reactor and the event loops both run in synchronous mode and can be treated in isolation, the overall result is a fully asynchronous service due to the decoupling of the reactor from the handled events. Annotation requests are added to all known (but not necessarily online) servers' queues (central and right branch of the reactor MAIN in Figure4.18); These might be new abstracts inserted into the queues manually for processing outside of high-load times (center) or stemming from a front-end request (right). Each server's loop iteration fetches an item from the server queue (starting with Server.run (thread)), resulting in various possible subroutine calls that can manipulate server and database state. Thread-locks are placed at key points (yellow stages shown in Figure 4.18) to avoid race conditions or dead-locks. Three objects model the entire back-end and are persisted in the BCMS database: a Server, a Queue (per server), and a Result (per response) object, with the corresponding methods described in Table 4.8.

4.4.4 Timing and Measurements

As the BioCreative II.5 challenge was held online via the BCMS with the already described modifications in place, the time taken by each server to annotate the test set could be measured. Figure 4.19 relates these measurements to the macro-averaged, document-centric evaluation results of each server. Pure article classification servers are not shown, as their time performance is generally below 30 sec/article. Due to the tech-

[28] A reactor pattern implements an event handler that concurrently responds to incoming requests and dispatches them synchronously to their associated "event loops". This pattern is found in asynchronous web servers, such as Python's Twisted, JavaScript's Node, or Ruby's EventMachine. It was first devised to serve advanced UNIX I/O polling implementations.

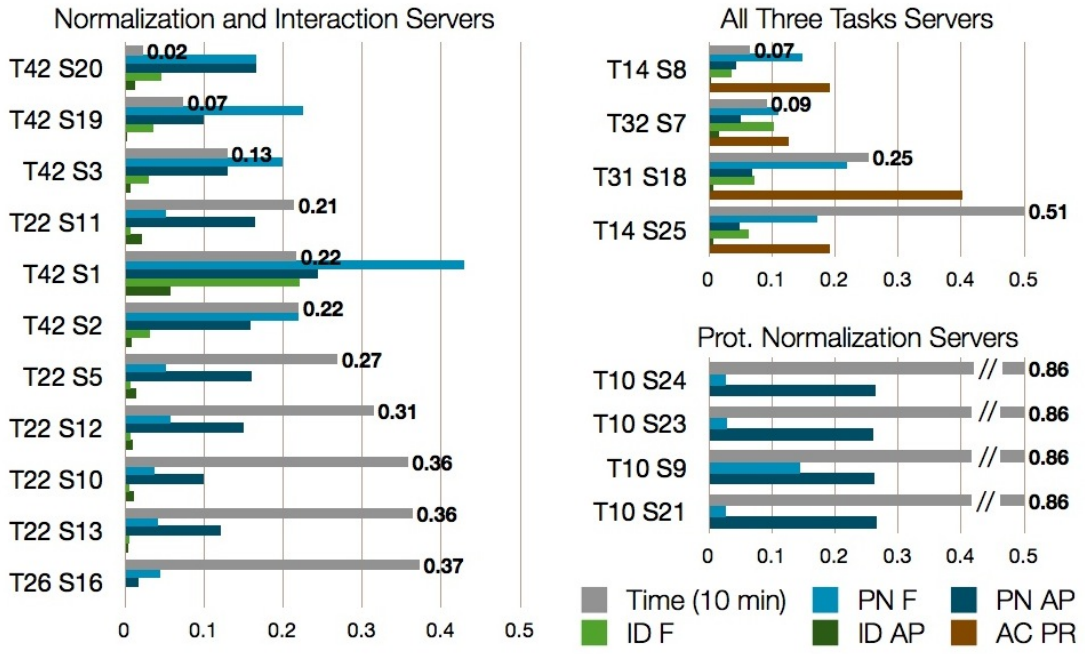

Figure 4.19: Timing and metrics of annotation servers during the BioCreative II.5 challenge. The time shown is the fraction of the 10 min a server had to annotate an article, i.e., a value of 0.1 equals one minute. The servers' document-centric evaluation results for the article classification (AC), protein normalization (PN) and/or interaction detection (ID) tasks are shown, ordered by time. Other abbreviations: F: (macro-averaged) F-score; AP: (macro-averaged) Average Precision; PR: AUC PR. Teams 14, 31, and 32 participated in all three BioCreative II.5 PPI tasks. The fastest server was S20 by team 42, taking no longer than 12 sec/article on average. The best protein normalization server by F-score was S01, team 42 (43%) - although the server only annotated 21/61 documents. The best normalization server with respect to AP was S21, team 10 (27%), annotating all 61 documents, although it was the slowest server, too (nearly 9 min/article). The best article classification server was S18, by team 31 (40% AUC PR; note that article classification only servers are not shown). The best interaction detection server was S01 by team 42 (22% F-score, 6% AP), although only 19/61 documents were annotated.

nical problems team 10 had during the test phase, their timing measurements should be considered outliers. In summary, the meta-server stored 1,404,709 annotations (i.e., after validation and dropping excess results) returned by the 27 participating servers[29] during the BioCreative II.5 *test* phase (approx. two weeks).

[29] split up between 8 article classification, 4 protein normalization, 11 normalization and interaction detection, and 4 servers that produced results for all three tasks

CHAPTER 5

Discussion

5.1 BioCreative PPI Corpora Discussion

The BioCreative PPI corpora form a gold standard for developing PPI-extraction systems. The articles were annotated by professional bio-curators following standard curation protocols established by the IMEx databases.

5.1.1 PPI Corpora Size

The BioCreative PPI corpora span 19,642 abstracts classified as PPI- relevant or not and 3,632 full-text articles annotated with experimental method and interaction pair data.

BioCreative and other related challenges[1] have consistently served the demand for creating sufficiently large collections of annotated text sources (corpora) that can be used for developing new methods by the text mining community. Indeed, most IE and IR methods are based on learning features from previously annotated text. Pioneering efforts in biology to create such corpora were started with the GENETAG corpus (Tanabe & Wilbur 2002), but also include the BioCreative collections, among many others[2]. To overcome the limitations of labor-intensive manual annotations, an interesting number of methods based on training with semi-annotated text have been proposed, too, the CALBC challenge (Rebholz-Schuhmann et al. 2011) being a prominent example. Still, sufficiently large collections with detailed annotations are a key limitation undermining the development of sophisticated text mining strategies. This makes biological corpora different from other NLP fields such as newswire, where these collections are readily available (see, e.g., (Sharoff 2006), for a review of "open source" corpora). In biology, both domain and expert knowledge are required, making it particularly challenging to create gold standard corpora for this field. Furthermore, the BioCreative collections are the only BioNLP corpora created in collaboration with professional database biocuration experts.

To our best knowledge, the BioCreative PPI corpora from the three challenges together form the largest gold standard for developing PPI-extraction text mining systems with the target of associating an article's content unambiguously with the relevant (protein) sequences, (protein) interactions, and experimental detection procedures. The corpora provide a broad basis for training and testing PPI systems for the challenges' tasks. They represent an independent, real-world standard for measuring and comparing system performance on each of these tasks. They are a lasting asset for the text mining

[1] see Section 1.2.1 for some examples
[2] see Introduction, Table 1.1 for some examples

community, since these freely available collections allow researchers to further improve systems and methods long after the challenges. Compared to collections of abstracts, a significant proportion of the content is available in full. Nonetheless, two issues need to be discussed: The first is related to the file formats, and the second to the annotations themselves.

5.1.2 File Format Issues

While PDF and HTML file extraction does not lead to the perfect plain-text generally found in BioNLP corpora, these formats represent data collections reflecting their real-world characteristics and availability.

The BioCreative II full-text articles are available both as HTML and PDF, and the BioCreative II.5 full-text articles are available as XML. As for these file formats, it has to be acknowledged that PDF files are not very suitable for text extraction (Ramakrishnan et al. 2010, Burns et al. 2003). The PDF extraction process does not always correctly reproduce the text flow, can lead to encoding errors for non-ASCII characters, or even terminate prematurely. However, the BioCreative III full-text articles were only provided as PDF files. The HTML files of BioCreative II were manually collected by the organizers and the XML files of II.5 were provided by the Elsevier publishing house. However, for BioCreative III, most publishers only granted the rights to redistribute the PDF files.

Having to use delicate syntactic parser on noisy text data will hamper the use of this advanced NLP technology. Preprocessing errors due to loss of text flow, dropped spaces, wrong or missing characters, etc. will have a detrimental effect on identifying word boundaries and text flow. This will negatively influences the performance of language processing tools. It can affect protein normalization, because protein NER relies on term boundaries, too. And encoding errors occur typically with non-ASCII characters such as greek letters. These in turn are significant markers to identify the correct mapping. For example, if the text span "SSRβ" (SwissProt:P43308) were garbled to "SSR?", even if a system were able to successfully identify "SSR?" as a protein mention, it likely would normalize preferentially to SSRα (SwissProt:P43307), not the beta-subunit. The effect of imperfect extractions is even worse for statistical machine learning frameworks, such as Markov chain-based systems, that rely on models created from (word) token sequences. If the correct sentence and token boundaries can no longer be detected (e.g., because spaces are dropped or lines get miss-split), their performance will decrease. This in turn means that establishing syntactic relationships between two protein entities mentioned

in a sentence will be impaired. Nonetheless, the difference in performance of syntactic parsers between full-text and abstract (see BioNLP Shared Task 2, (Kim et al. 2011)) is possibly based in the difference of language use between a publication's body and its abstract, too. Furthermore, parser performance is inverse to sentence length (Verspoor et al. 2009).

While "perfect" text sources might result in increased performance of mining approaches, the ability of working with noisy input is one of the main issues in data mining. It should rather be accepted as a challenge in scientific studies, not avoided or ignored.

5.1.3 Document-level Annotations

While the annotations on the corpora are mostly only at document level, MINT and IntAct curators provided evidence passages for the interaction pairs and experimental methods of the 338 full-text test set articles from BC II.

Second, the annotations on the corpora are (mostly, see below) provided at document level. In other words, the text passages that lead to the annotations made by the PPI database curators are not recorded. In this respect, the BioCreative II PPI test set annotations form a unique exception, where MINT and IntAct professionals curated the passages that they used as evidence for their annotations. For text mining systems, the ability to learn from the exact expression or phrase that gave rise to an annotation is incredibly valuable for training and developing classifiers.

For example, the following figure caption[3] probably was sufficient for the MINT curators to annotate four cross-species PPIs interacting by co-localization with fluorescence microscopy-based experimental evidence:

> **Immunofluorescence analysis** of stable **C2C12 cells** expressing **human affixin**. [*...other sentences...*] Affixin **co-localizes** with ILK, αPIX, βPIX and dysferlin lamellipodium tips (arrowheads) in the **C2C12-affixin cells**. Scale bar, 20 μm.

The first sentence indicates that "immunofluorescence analysis" is one of the experimental methods described in this publication. The last sentence provides evidence that the co-localizations of affixin with ILK, αPIX, βPIX and dysferlin are shown. The first sentence also contains important clues for the protein normalization. Affixin was cloned

[3] Taken from a BC II.5 PPI article, DOI: 10.1016/j.febslet.2008.01.064, figure 1

from H. sapiens, while the presence of "C2C12 cells" indicates that all other protein names should be normalized to their mouse proteins.

If the curators' annotations were associated to these two sentences, text mining systems might be able to "learn" from these associations. For example, in the above excerpt, the verb "co-localize" could be detected as a marker (commonly called *interaction verbs*) for mining PPIs from sentences containing it, and that "immunofluorescence analysis" is an indicative term for annotating the correct PSI-MI ontology interaction method *fluorescence microscopy* (PSI-MI ID MI:0416).

While not all articles are covered this way, for 338 of the test set full-text articles in BioCreative II, the MINT and IntAct curators provided the evidence sentences that they had judged as the most informative sentences for their annotated interactions.

5.1.4 Inter-Annotator Agreement

The high IAA is an excellent indicator for the quality of this gold standard. The lower inter-DB agreement on pairs reflects the increased difficulty of the interaction detection task.

The intra-database IAA scores for protein normalization (Table 4.3) indicate the true, expected annotation consistency and match well with the F_1-measures of the curator evaluation (see Author-Curator-System Comparison, Figures 4.14 and 4.15). However, the significantly lower **inter**-DB IAA (76%, $\kappa = 0.36$) for the interaction (pairs) annotations shows that the different curation protocols used by the databases results in disagreements on interaction annotations. Furthermore, these cases of disagreement are representative for the increased difficulty of the interaction detection task.

This agreement study should not be confused with an evaluation of IMEx database quality, which has been under significant scrutiny recently (Cusick et al. 2009). The overall IMEx annotation consistency has been shown to be generally higher than the results in this work or the cited criticism; see (Salwinski et al. 2009). The IAA data here is used to express the annotation consistency on the BioCreative corpus, not the quality of the IMEx curation process.

5.2 Evaluation of Ranked Results

One objective of this work was to explore suitable evaluation metrics for measuring the quality of the ranked results submitted by the participants.

5.2.1 Conditions for Evaluation Measures

The ranking metric for the BioCreative results should reflect the total quality of the ***entire*** *annotations made by the systems.*

An ideal evaluation measure, as defined by (Swets 1969), should satisfy the following conditions:

1. It should express solely the ability of separating relevant from non-relevant results and not the method's efficiency or cost.
2. It should not be confounded by the relative willingness of the system to emit results, such as an "acceptance criterion", whether this criterion is a characteristic of the system or adjusted by the user. (In short, it should not depend on a threshold.)
3. It should return a single number so that it could be transmitted simply and immediately.
4. It should allow for complete ordering of different performances, indicating the amount of difference between systems in absolute terms. That is, the metric would scale with a unit, a true zero, and a maximum value.

Depending on the chosen school of thought, (Wilbur 1992) suggested the following modification to the second condition:

2. It should *be characterized by a threshold*, but should *reflect the quality of retrieval at every rank down to that threshold.*

The main argument of this modification's proponents is that not applying a cutoff can lead to a classifier evolving towards disproportionate result sets because of small performance increases provided by low-ranked hits inherent to most ranking quality metrics. As can be seen from the AP results, this is especially important if the number of classes (to annotate) is small[4].

The main counter-argument to this modification is that classifiers would be tuned towards that (subjective) threshold. Instead, they should be trained and evaluated by using a measure with a universal optimization condition.

An important difference implied by Wilbur's modification is that no longer the entire result set is evaluated, but only the results before (or, up to) the threshold. I.e., while systems no longer gain performance from disproportionately increasing result sizes, no penalty is applied to any present excess results, either.

[4] As, for example, in the case of the method extraction task, where the FAP-score based cutoff did change the rank order of team 65 w.r.t. AP (see Section 4.3.4).

5.2.2 The FAP-Score Measure

The FAP-score is an evaluation metric that fulfills all of Swets' conditions for evaluation measures.

The most common metric for measuring ranked result sets with multiple, relevant hits are probably the AP (that is a variation on AUC PR) and AUC ROC measures. Between the AUC of the ROC versus the PR curve, it has been shown that a curve dominates in ROC space if and only if it dominates in PR space (Davis & Goadrich 2006), especially on skewed datasets, such as for protein/gene normalization tasks (While there are virtually millions of false annotations - i.e., UniProt accessions - only a few are true.)

However, as can be seen from the results presented for the protein normalization, interaction detection, and method extraction tasks, the AP values can be inflated by generating large result lists. Recently, a thresholded variant of the AP measure (TAP_k) was introduced, circumventing this shortcoming (Carroll et al. 2010) (see Section 3.2.3, too). However, as common with thresholds, it is impossible to define an optimal cutoff. The cutoff will depend on personal preference, while at the same time system-specific behavior and parameters will result in different ideal cutoffs. This is can be seen by example from the specific "ideal" cutoffs reported for systems in the Results – each system, run, and task has been shown to have different a threshold. As TAP_k conforms with the modification introduced by Wilbur, systems are not penalized for irrelevant results reported after the threshold either.

By combining F-score and AP (see Materials and Methods, Section 3.2.5), we have experimented with an non-thresholded rank quality metric that can be used

- to identify the best cutoff by determining the highest FAP-score in an unbound ranked list and
- to express the performance of a system that produces a (bound) ranked list *without* a threshold and as a single, comparable number.

In other words, the FAP-score fulfills all the conditions suggested by Swets, because it does not require defining a threshold, while it reflects the quality of retrieval for the entire result set. The FAP-score penalizes disproportionate result sets, because it evaluates the entire list, not just part of it.

The most important property of the FAP-score is that is formed by the harmonic mean between the F-measure and AP. From the three Pythagorean means (arithmetic, geometric, and harmonic), it always produces the least mean. This property is the main

reason why it is possible to find an optimal cutoff for unbound ranked lists. Were the other two means applied to calculate the FAP-score at every rank in a result list, several different ranks would produce similar maximum results. Such a behavior would resemble a violation of Swet's first and fourth criteria, and make it impossible to determine the best cutoff position in unbound lists or compare the quality of systems producing bound lists.

5.3 Discussion of Evaluation Results

5.3.1 Article Classification Systems

Three of the four best classifiers were mainly SVM-based. The full-text articles were more difficult to classify than abstracts.

Examining the evaluation results, the BioCreative II ranking approaches (AUC PR) seem to be significantly better than those of II.5 or III. However, the balanced situation of positive and negative articles in the BC II test set has less impact on precision, explaining the systems' seemingly better PR curves when compared to the curves of systems in the other two challenges. If the PR curve of team 6 shown in Figure 4.4 is to be directly compared to the other teams, it is necessary to take the different article distributions (between relevant and irrelevant articles) into account. This effect can be *roughly* estimated by stretching the team's curve downwards until it meets the same precision at full recall (i.e., the point where the curve hits the right border of the plot) as the curves of the other teams (i.e., from 50% down to about 15% precision). With this distortion, the curve of team 6 would not appear that different from team 20, BC II.5.

The classifier used by three out of the four best performing teams[5] was SVM-based. Only team 9 (BC II.5) used their own Variable Trigonometric Threshold (VTT) classifier, which uses a linear decision surface based on the cosine vector similarity between features for separating the articles. Interestingly this "classical" cosine vector similarity separator achieved the best MCC score (0.58), ahead of the second best results (MCC = 0.55). Team 6 in BioCreative II used a linear kernel, while team 28 explored Naïve Bayes, Max. Entropy, and SVM classifiers. They tried using both a linear and a string (p-spectrum) kernel, the latter outperforming all other approaches. Team 73 (BC III), too, used a linear SVM, but replaced the regular hinge loss with the Huber function.

[5]defined by their MCC scores

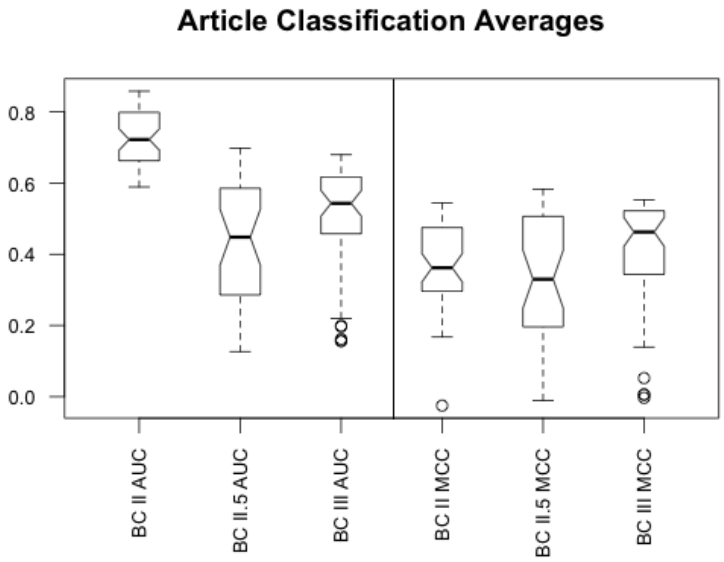

Figure 5.1: Box plots of the average AUC PR (left) and MCC scores (right) in the three PPI challenges.

BioCreative II and III participants had only abstracts to classify the articles and the publications are taken from a mixed set of journals. During BioCreative II.5, participants received the full text of the articles and the articles were all from FEBS Letters only. Furthermore, while team 9 (BC II.5, with the highest MCC score in all challenges) also participated in BC III (as team 81), they reported that a software error might have been the main reason for a significantly lower performance in the third iteration of the task. The overall classification performance of all participants generally improved from II to III, as can be seen from the MCC averages shown as box plots in Figure 5.1 (right three plots). In other words, an overall increase of the field on this particular task can be observed. Furthermore, the more spread results of BC II.5 (compared to II and III) could be interpreted as an indicator that full-text classification is more difficult than abstract-based classification.

The differences in average AUC PR from BC II to the other two challenges shown in Figure 5.1 stem from the different balance between true and false articles in the test set. BC II articles had a balanced distribution, while in II.5 and III the test sets only contained about 15% positive instances. This can be seen by comparing the PR curves of the systems, as explained before.

System Comparisons

Four features could be identified as highly relevant to the task. Creating a balanced training set has been shown to improve classifier performance.

The best systems have several features in common. It is apparent that word (token) uni- and bigrams form a good baseline feature set. Furthermore, character n-grams (of length 6 or 7) were found to be another useful baseline feature. These bag-of-word feature types need to be filtered and/or weighted for relevancy, where it seems that the information gain and chi-square statistic were the most successful feature selection strategy. While this observation is in concert with known feature selection strategies (Yang & Pedersen 1997), this approach was not reported by all teams[6]. Team 9 used only one feature in addition to n-grams, while achieving similar performance as the other three teams. This fourth feature was generated from a gene/protein mention recognition approach. All four teams (and several others, too) made use of it. Here, it was shown by the participants that simply counting the number of detected gene/protein mentions in the abstract (and figure captions, if full-text is classified) is sufficient. In other words, the protein/gene mentions themselves are not necessarily features that lead to a good separation.

These four features (word unigrams, word bigrams or pairs, character 6- or 7-grams, and a gene/protein mention count from the abstract/figures) combined with a feature selection strategy are the common grounds for good PPI article classifiers. Finally, training on a balanced set[7] using over- or under-sampling was shown to have a positive impact on performance by several of the participants independently (Kolchinsky et al. 2010, Lan & Su 2010, Schneider et al. 2011, Lourenco et al. 2011).

5.3.2 Protein Normalization Systems

Protein normalization systems achieved an F-score of 29%, and an AP of 26%. The FAP-score could determine the best cutoffs for teams 10 and 22, identify the best normalization submission (10-S09), and produced a reasonable ordering of the team results.

The best protein normalization systems during BioCreative were provided by teams 40[8] and 42[9] if measured by F-score. However, the six best FAP-score based runs are: 1.

[6]However, these feature selection strategies were present in the system descriptions of the four best teams.

[7]as the BioCreative article corpora training sets are non-balanced

[8]BC II, macro-averaged F-score = 29%

[9]BC II.4, document-centric, macro-averaged F-score = 56%

10-S09 (19%); 2. 18-5 (18%); 2. 37-3, 42-S02, and 14-3 (all 16%); 4. 22-3 (9%). While we have to acknowledge that we did not provide any (mathematical) proof that the FAP-score ordering is superior to the AP or F-score ranking, this ordering makes intuitively sense (see Figure 4.6). And, as can be seen from the initial results, the probably most powerful normalization system was provided by team 10, if measured by AP. However, in AP ranking, the result of team 22 would be second best (see 4.3.2), while in F-measure it would be very low ranked; Both these relative orderings could be misleading. And, although team 18 (run 5) has the best F-measure, their FAP-score is only 18%. This is reasonable, too, due to their comparatively poor ranking performance (AP = 14%). Furthermore, as we will show later, picking these six runs and combining them using the FAP-score and ensemble voting can produce an astonishingly good overall result (see Section 5.3.3).

By applying the FAP-score based cutoff established on their training data, 10-S09 can (theoretically) even achieve a higher F-score (39%) than any other team in both challenges. And team 22 might have achieved an F-score of 30%, too. Nonetheless, the new cutoffs would not have changed any of the three high-recall teams' (including team 51) positions in the team ranking if ordered by AP. It can be stated that the attempt to boost AP is mostly noise appended to otherwise excellent result sets. Furthermore, using a cutoff would have put especially team 22 well ahead of other teams in terms of F- and FAP-score.

System Comparisions

NER of gene/protein names, gene/protein name abbreviation detection, as well as their regularization facilitate normalization approaches. Context information is supplied to a classifier during the matching and/or disambiguation steps and increases performance.

All normalization systems follow three process stages (regularization → matching → disambiguation), and many teams add an initial NER step, including the four best teams. However, at all of these stages nearly every team's particular approach is distinct. It is non-trivial to identify the most important features that form the basis for an "ideal" approach.

An advantage of using NER as the first step is that regularization and/or approximate string matching can be limited to the detected names only. Nearly all teams regularize the text and/or regularize the content of their dictionaries before the dictionary matching step. Without NER, most teams report the use of exact matching approaches

because of the better precision, and proper regularization should render the need of an approximate matcher mostly obsolete. An exact, but case-insensitive, partial matcher against regularized NER terms seems to be one of the more successful strategies, but this is largely dictionary-dependent. For example, gene/protein names are sometimes tagged with affixes referring to experimental or organism context, e.g. “anti-E2F”, “hHR23A”, “BRCA1-luc”, “pB-junB-infected”, or more obscure cases such as “pDB7JunD/eb1”, describing a vector containing JunD. In these cases, substring matches of the mention with the dictionary entires have to be detected, explaining why partial matching can be useful. And even for approximate matching, there is not sufficient evidence to principally discourage any fuzzy matching approach. In general, it might be interesting to attempt a mixed approach (i.e., exact matches for symbols, partial matches for “clean” names, and fuzzy matching for long, “artificial” database names containing commas, parenthesis, etc.).

Another common property of many normalization approaches are abbreviation handling capabilities. A dictionary might contain the full name of a protein, while authors use an abbreviated form of the name (not present in the dictionary), or vice versa. A few systems are able to detect these author-introduced abbreviations for full protein/gene names (e.g., team 42, BC II) or keep track of the uses of both the full name and (abbreviated) symbols (e.g., team 6, BC II). By keeping track of the various normalizations of the synonyms, it can be easier to identify the correct normalization for all the variants of a protein name present in the text.

The last distinctive property we will discuss is the use of context information. Capabilities to detect organism mentions are ubiquitous in all but one system (from BC II), but not all systems include names of cell lines and their corresponding species mapping. Prefixes are often used describe the organism source (h, m, r, v, At, etc.) and some systems detect them. The more commonly used context information beyond species mentions were sequence length, genomic location, and molecular weight. However, a few systems make use of available concept annotations on genes, such as GeneRIFs or GO Annotations (e.g., team 10). Quite a few teams examine the sentence of a mention to detect formulations that are used to describe protein interactions, but other teams leave this issue to the interaction detection step. These patterns might include keywords that can indicate experimental procedures (such as “co-localize”, “precipitated”, etc.).

One technique not seen so far that might improve the systems is the use of OCR (optical character recognition) on the figures in the publications. Authors often tend

to add arrows with the names of proteins to gel or fluorescence images. DB curators have named figures as central elements in the curation process, both for deciding on "curate-ability" and for locating the relevant information[10]. While in nearly all cases these names will be mentioned in the figure captions and the body text, too, it might help identify the most relevant proteins implied in experimental methods. However, it is not clear if current OCR technology is able to extract these names with a sufficiently high precision. And even if the quality were sufficient, it is neither obvious that this information would lead to increased performance.

Normalization System Errors

The four major issues current systems need to overcome are (1) improving organism disambiguation (i.e., assigning the protein mentions to the correct species), (2) expanding the dictionaries with protein/gene names from species-specific resources, (3) extracting proteins with experimental evidence only, and (4) detecting long range (discourse) relationships.

The major aspects of normalization issues can be grouped by error type. Errors stem from incorrect organism disambiguation and from normalizing proteins that have no experimental evidence[11]. False negatives on the other hand stem from incomplete dictionaries and discourse (only) relationships[12].

The **organism disambiguation** issues are mostly related to the failure of creating the correct associations between organism mentions and the relevant proteins. This is especially relevant for homonymous names used to identify proteins/genes in multiple species. For example, for the figure caption example shown in Section 5.1.3, systems would have to be able to reliably detect that affixin and only affixin should be normalized to a human protein ID, while all other mentions must be normalized to mouse (and not human).

We also took the organism-specific performance of the six best runs from the six best teams in both challenges. We split the results by organism, grouping the annotations

[10] Personal communications with bio-curators at various conferences.

[11] FPs from normalizing proteins that are not part of an interaction are present, too. But these FPs appear to have had a lower impact, especially with more advanced systems that prioritize proteins for which they can also detect an interaction-describing sentence. In other words, we could not find sufficient evidence that would prove this is one of the major errors.

[12] These occur if an author introduces a protein in a sentences and then refers to it with a pronoun in the following sentences. At the same time, the following sentence describes an interaction with another protein. While it is very difficult to quantify the discourse relationship error, 5% of all interacting protein pairs are not co-mentioned in a single sentence. But the more frequent discourse relationship errors are between interacting proteins and their experimental evidence or organism source.

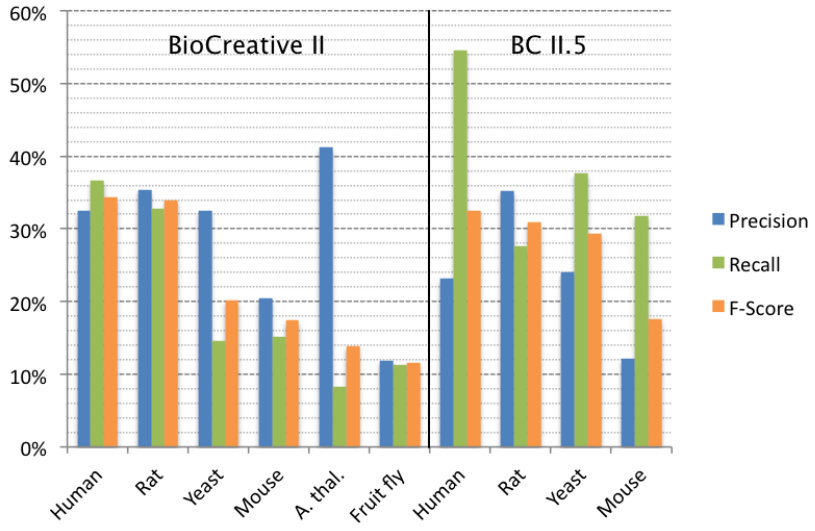

Figure 5.2: Organism-specific influence on the protein normalization performance of the six best systems of BioCreative II and II.5. BC II runs: 6-4, 17-3, 19-2, 28-1, 40-2, 42-3; BC II.5 runs: 10-S09c7, 14-3, 18-5, 22-3c9, 37-3, 42-S02. A. thal.: thale cress (Arabidopsis thaliana)

from all runs per species, normalized them[13], and calculated the micro-averaged precision, recall, and F-score for each species. This produces the histogram shown in Figure 5.2. The analysis clearly shows that mouse proteins were almost as difficult to normalize as Arabidopsis and fly proteins, while human and rat proteins were substantially simpler to disambiguate. The mouse issue is in part explained by publications referring to the human proteins when describing murine interactions. A more obvious explanation is that mouse and human proteins and genes have many names and symbols in common, making the disambiguation issue more prominent. And, many systems use human as the default organism. This explains the better recall over precision for mouse proteins in BC II.5, as shown in the histogram. The inter-species ambiguity likely explains why rat proteins had better performance, too: They do not share as many homonyms as M. musculus does with the human nomenclature.

As can be seen from the organism filtering results, the FP error introduced by normalizing to irrelevant organism more than halves the systems' performance. And, this negative impact does not include the FN errors from missing out on relevant organisms, or FP errors, especially when multiple species mentions occur in an article. It therefore can be safely stated that organism disambiguation to date is the most prevalent

[13] I.e., we summed up all TP, FP and FN annotations per species from the six runs and divided the sum by six.

normalization error.

As for **protein dictionary** issues, the systems missed names that are not present in cross-species repositories such as UniProt or Entrez Genes. This points to investigating the appropriate use of additional, species-specific databases such as HGNC, RGD, TAIR, FlyBase, etc.. It is probably the main reason for the low recall of Drosophila and Arabidopsis proteins in Figure 5.2. For Drosophila, this might be coupled to the fact that fly names are sometimes homonyms of regular English words, and some teams filter those homonyms from their dictionaries.

Regarding the impact of the dictionary issues, it is impossible to quantify this error with certainty[14]. But an advantage is that this problem is straightforward to lessen: Systems will need to explore ways to expand their dictionaries. We even found examples of cases where the UniProt names were incomplete for human proteins[15], not just fly or thale cress. In general, these cases can be traced for proteins that were missed by all or most systems.

Experimental evidence detection failures mostly contribute to FPs. One problem is that there is no sufficiently large corpus available with expressions referring to annotated experimental methods. With such a corpus, a NER system could be trained to improve the detection of associations between proteins and experimental methods. However, such expressions would at least have to be grouped by the applicability of the methods, e.g., for genetic, protein-protein, protein-DNA, and metabolite interactions. Second, experimental evidence and interaction evidence is often separated. For example, on a corpus of 62 full-text articles we are currently curating (unpublished data), we found 1167 interaction sentences, 2868 experimental evidence sentences with one interacting entity mentioned only, but only 730 sentences where both entities and the experimental evidence are mentioned. Experimental evidence detection might be improved with an already discussed image OCR strategy.

Finally, the **discourse relationships** issues arise with both organism disambiguation and experimental evidence, as well as with detecting protein interactions themselves. As for the experimental evidence and/or species, they are often mentioned in another place than the interaction. These kind of relationships will be very hard to detect,

[14]It is clearly lower than organism disambiguation, but this would depend on the distribution of articles. Our organism distribution in the gold standards is intended to reflect an average organism distribution across PPI articles.

[15]I.e., names present in HGNC, but not UniProt, such as *ubiquitin-activating enzyme E1* for UBA3 (UniProt:Q8TBC4).

at least with current approaches: It can be stated[16] that all systems present in the challenge essentially are sentence-based approaches. Furthermore, detecting discourse relationships and discourse analysis are generally accepted as one the most difficult tasks in NLP.

By "discourse interactions using experimental evidence" we refer to cases where a protein is named in an experiment, but not explicitly described in an interaction. This is sufficient evidence for DB curation, and thus part of the gold standard. The paper might describe an experimental procedure used on one of the main proteins. Then, the authors go on to describe that they detected a list of proteins with this method, possibly sentences or even sections apart. All these proteins will form part of the gold standard, but are nearly impossible to detect with the current approaches.

An example of such a case is PMID 16712842 (DOI: 10.1016/j.febslet.2006.05.012), where mKIAA0467 is shown to interact with hHR23A via immunoprecipitation. However, mKIAA0467 is only mentioned once, in a list of proteins found on a gel. From the discourse it can be understood that this is the proof that mKIAA0467 interacts with hHR23A, while hHR23A is not mentioned in the entire paragraph where mKIAA0467 appears. Therefore, the hardest error to overcome will be the ability to find long-range discourse interactions and relationships. And, this problem is even more difficult to solve when protein normalization is coupled to requiring experimental evidence, too.

Another problem, that can be seen in the same article, is the difficulty of assigning organism mentions to proteins (i.e., species disambiguation). Here, authors refer to two homologous proteins from two different species using the same name[17]. In the figure caption example shown in Section 5.1.3, this issue can be seen, too: while the sentence states "C2C12-affixin cells", particularly affixin may not be normalized to its mouse protein. The clear explanation of the relationship of affixin to human is only found a few sentences earlier.

SwissProt- and UniProt-based Normalization

The UniProt knowledge base is a better source for gene/protein dictionaries than SwissProt only, despite the lower curation standards of the TrEMBL subset.

A possible issue in the assembly of protein normalization dictionaries is the question

[16] Team 6, BC II, had some very minimal discourse detection strategy. But even they concluded that it was too noisy to be truly useful.

[17] yeast Rad23 and human HR23A/B, while using Rad23 interchangeably to refer to both the yeast or human proteins

of using SwissProt data only or integrating TrEMBL data, too (i.e., using the entire UniProt database). Because most relevant proteins that had to be normalized are part of SwissProt, having more ambiguity and possibly adding protein names not even used in literature from the much less curated TrEMBL records might be undesirable. This gives reason to argue that using SwissProt only for the normalization task might be sufficient and even beneficial.

The protein normalization tasks asked participants to normalize the protein IDs to SwissProt and UniProt accessions for BioCreative II and II.5, respectively, while both sets have SwissProt and TrEMBL IDs in the gold standard. Therefore, the difference of using the two resources for normalization could be compared here. The results showed that if evaluating against SwissProt only, the average results increase 3.3pp for BC II and 2.1pp for II.5, as would be expected. However, if removing all TrEMBL accessions from the BioCreative II.5 system results, the average performance actually degrades by 1.6pp. Furthermore, the best scoring teams' (10, 18, 22) scores degraded between 3pp and 5.4pp or nearly up to 20%.

While this cannot be claimed as a final proof, the results do clearly indicate that the better the system, the more likely it will benefit from being able to choose from the full namespace for their normalizations. A UniProt-based dictionary seems to have been an advantage for these systems, even if only 7% of all normalizations are for TrEMBL accessions, as in the case of the BioCreative II.5 gold standard. Indeed, manually investigating protein normaliziations missed by all systems showed that systems should rather go beyond UniProt as resource for their dictionaries, and only limit it at the species level. In most cases, the author-used name or at least a very similar lexical variant could be found in the specific database for the target organisms[18]. However, resolving to these names/including them in the dictionary will increase name ambiguity, so it will require further research to quantify how much can be gained.

5.3.3 Ensemble Normalization Comparison

An ensemble annotation could potentially increase the result quality. The result indicates that there is room for improving the systems participating in the challenge. The FAP-score was instrumental to produce a better ensemble result.

In protein structure prediction, ensemble systems (i.e., protein structure prediction meta-servers) have been shown to outperform single systems (Valencia 2003). This same

[18]We searched HGNC for the human, FlyBase for the fly, MGD for the mouse, YGD for the yeast, TAIR for thale cress, and RGD for the rat proteins that had been missed by all systems.

question has been investigated for the annotations mined by text mining systems. Here we wish to investigate the theoretical performance an ensemble result can achieve if modeled from the protein normalization results.

A simple approach was used to combine the top six runs of BC II.5 (10-S09, 14-3, 18-5, 22-3, 37-3, 42-S02)[19]. For all runs, cutoffs were established on the *training* data, as the confidence value that produced the best FAP-score[20]. We then combine all runs by voting, requiring at least three of these systems to have reported the same annotation. This produces a result set that has a macro-averaged precision of 49%, a recall of 45%, and an F-measure of 44%. It provides annotations for 55 out of 61 articles, reporting on average 3.5 annotations for each of those 55 articles (I.e., the ensemble result provides no annotations for six articles.)

Concerning the use of the cutoff runs in the ensemble, it needs to be clarified that these values were established without using the test data (see Materials and Methods, Section 3.3.3). We "predicted" the cutoffs on the training data and these confidence value cutoffs were applied unaltered on the test set results. We therefore can argue that it is possible to treat the ensemble procedure as plausible.

Another two issues are how much the ensemble result improves over the (cutoff) runs and how much this procedure contributed to improving the ensemble result. The test set result of 10-S09, limited to annotations with 0.04 confidence or better, achieves a macro-averaged F-measure of 39%[21]. So the ensemble is 5pp better, for an increase of 13% over the best run it contains. If creating an ensemble result from the runs without applying any cutoff, the F-measure of the ensemble would be 38%[22]. This difference shows that the FAP-score can be applied to create a 6pp or about 16% better ensemble result.

The ensemble indicates that it could improve the results generated by the BioCreative protein normalization systems. The ensemble result furthermore indicates that the individual result sets are heterogeneous and participants are not using the same methodology. If all systems had annotated the same proteins, the ensemble result would have

[19]These systems and runs achieved (macro-averaged) F-scores/APs of 15/26, 22/13, 27/14, 5.4/23, 23/13, and 20/13%, as reported in the Results.

[20]This is achieved by searching for the minimum confidence value to use that produces the highest FAP-score on the run's training data, in 0.01 increments. Runs 42-S02 and 18-5 had no such cutoff, because the results of team 42 seem to be optimally constrained already, and team 18 reported a confidence of 1.0 for all results. For the other runs, the confidence values were 10-S09: 0.04; 14-3: 0.57; 22-3: 0.71; 37-3: 0.07.

[21]prec. = 39%, recall = 46%

[22]prec. = 33% precsion, recall = 57%

been significantly lower. In other words, it shows that no system alone is yet "perfect" (even though 10-S09 is close) and there are possibilities for improvements by learning from each approach.

We also measured an ensemble result using the document-centric evaluation with organism filtering (see Materials and Methods, Section 3.3.3). Given that organism filtering increases precision, it is better to make use of all annotations that are provided by two (instead of three) of the systems. This change will increase recall at the cost of precision. This organism-filtered ensemble provides annotations for 54 articles (89%). The filtering procedure is able to eliminate half of the remaining false positives in the ensemble. The macro-averaged F-measure for the ensemble result is 57%[23].

5.3.4 Interaction Detection Systems

The main issue of PPI pair extraction is the normalization error. BioNLP systems have performance results in the approximate range of 40-60% F-measure if PPI extraction is treated in isolation (i.e., ignoring entity recognition and normalization). However, in the challenges' settings the propagation of normalization errors (< 30% F_1) resulted in F-measures below 15%.

The performance of text mining systems extracting PPI pairs is low, with the best F-measures not even surpassing 15%. Reporting pairs in a simple, combinatorial approach takes the n protein normalization results and outputs the $\binom{n}{2}$ (or $\binom{n}{2} + n$ if allowing for auto-interactions, too) possible ID pairings. This combinatorial approach would have the best possible recall, but the lowest possible precision. The challenge in this task is finding the true associations within all these possible pairings, and detecting the corresponding, experimentally backed interaction evidence in the text.

We can compare the scores of the document-centric, organism filtering evaluation of the interaction detection results to their raw scores (Figure 4.8). The systems magnify the species disambiguation error introduced by the normalization systems (see Figures 4.13 and 4.5). The F-measure of the normalization results increases by an average factor of 2.2 after organism filtering, while the average factor is 4.8 for interaction detection[24]. I.e., the species disambiguation errors undergo a combinatorial boost. As a matter of fact, the document-centric, organism-filtered *interaction* result of team 18 even outperforms the team's raw *normalization* result. If text mining should more reliably extract protein or genomic interactions (e.g., gene regulation) mapped to their DB sequences

[23] prec. = 58%, rec. = 65%, see Figure 5.5

[24] using each team's best raw run, and none of the cutoff runs

from full-text articles, the performance of current normalization approaches will have to improve substantially.

In general, detecting interactions depends on the interaction type. For example, systems are quite capable at detecting PPI phosphorylation events from text. According to the BioNLP 2011 Shared Task results (Kim et al. 2011) (see "EPI task" results), at an F-measure of 83%. Usually, results have a higher precision than recall, with F-scores in the 30-80% range. The scores vary with the particular event type. The worst interaction event type in the EPI task was DNA demethylation, with an F-score of 27%. The "traditional" task in that challenge is the GENIA task. It evaluates general event extraction performance. In particular, the "GENIA task 2, sites only" results evaluate associations between two entities and the relevant event type (e.g., `phosphorylation(PDPK1 → Akt1)`). The best score of a participant team on reporting the correct associations between two entities and the relevant event type is 38% F_1 on full-text articles at 58% precision (Team UTurku). In this BioNLP challenge, too, the differences between full-text and abstracts are apparent: On the abstracts, this team reached an F-score of 57% for the same task. If split into the three main event types (phosphorylation, binding, and regulation events), there are huge differences between the events and teams. This means that different approaches do better depending on the event type. The best phosphorylation extraction approach achieves an F-score of 84%, but general binding and regulation extraction only reach 37% and 41% F_1, respectively.

Another independent study examined the performance of kernel-based methods to extract PPIs (Tikk et al. 2010). They tested approaches relying on rules using shallow syntactic information and patterns, phrase structure, and dependency trees over five different corpora (AIMed, BioInfer, HPRD50, IEPA, and LLL) and with blinded entities[25] Therefore, neither NER or normalization performance influence the results. They then measure interaction detection performance of the approaches when trained on the same corpora as they are evaluated on, as well as measuring the approaches when tested on another corpus. Similar F-scores were measured as those reported by the BioNLP 2011 Shared Task. Noteworthy evaluation results are the scores from the larger AIMed and BioInfer corpora[26]. Kernels using the phrase structure performed worst, while only the dependency tree-based APG kernel (Airola et al. 2008) was able to slightly outperform

[25] "Blinded entities" means that all mentioned proteins were manually marked so that the system could immediately recognize them.

[26] Those two corpora have a higher proportion of sentences containing two or more proteins with no actual interactions than the other three. In addition, they contain far more sentences than the other three.

the rule-based shallow syntactic kernel if evaluated on the same corpus. The APG kernel achieved 61% F_1, while the shallow, pattern-based approach reached 60%. However, in the cross-corpora evaluations between AIMed and BioInfer, the shallow linguistic approach did not drop below 41% F-measure, while the worst APG kernel performance was 23%. Tikk et al. concluded that shallow linguistic features and patterns worked more reliable than using (deep) parsing approaches in the cross-corpora settings, and that the shallow approach had performance equal to dependency parsers in the intra-corpus evaluation.

To compare these results with the BioCreative interaction tasks, it is necessary to be aware of the main difference of the BC tasks to the BioNLP Shared Task evaluation and the evaluations by Tikk et al.. The latter two measure the semantic annotation quality at the sentence level, while the BioCreative annotations are measured at the document level. The important interactions can be expected to have been mentioned multiple times within a paper by the authors. This should increase the performance at document-level. At the same time, (Alex, Grover, Haddow, Kabadjov, Klein, Matthews, Tobin & Wang 2008) report that 5% of all interaction pairs are not co-mentioned in a sentence. Furthermore, it can be assumed that not all relevant, co-mentioned pairs will be described as interacting in at least one sentence. Taking these issues into account, it seems appropriate to say that the two evaluation results presented above can be used as (very) rough estimates for a raw document-level interaction detection performance of current systems (i.e., the performance without NER and normalization issues). However, with less than 30% normalization F-scores, normalization performance is generally lower than (raw or blinded) interaction detection. We have already shown the combinatorial "boost" of normalization errors in the interaction detection results, too. Therefore, we conclude that the larger (normalized) interaction detection error is introduced by the protein normalization.

System Comparisions

Systems relying on shallow parsing and interaction patterns can be competitive to deep parsing approaches.

Already during BioCreative II it became apparent that pattern-matching approaches as well as shallow or deep parsing had an advantage over purely statistical NLP methods. This likely explains why the systems in BC II.5 were all based on at least one of the former three, with team 22 even integrating them all.

It should be noted that only two teams in the two challenges used dependency parsers (40, II and 22, II.5) and team 22 in addition used their phrase-structure parser, Enju (Sagae et al. 2007). All other teams in both challenges only reported the use of shallow parsers (chunkers)[27], and/or pattern matching. However, three other teams (6, 14, and 18) were able to achieve similar results to teams 40 and 22 (see Figure 4.8). The shallow parsing and interaction pattern matching approaches of those three teams could mimic the performance of the deep parsers (that also used patterns). And, run 18-5 had the best F-measure of all five. This result is mostly in accordance with the comparison of kernel-based PPI detection approaches described by Tikk et al.. However, it has to be acknowledged that the only two teams using deep parsing approaches were both in the league of the best performing systems.

From Abstracts to Full-text

The participant team relying on abstracts was able to achieve the same performance as the best performing systems using full-text articles. However, this abstract-based system is probably much closer to its maximum performance than the full-text systems, because the majority of relevant proteins and interactions are not mentioned in the abstracts.

Team 40 in BioCreative II processed the abstracts from PubMed only instead of using the full-text for the normalization and interaction tasks. This system is only overshadowed in terms of *protein normalization* F-measures by the full-text based systems of teams 10 and 22 in BioCreative II.5 *after* applying a FAP cutoff. In terms of interaction detection performance, only team 18 had a better F-measure. These result indicates that the negative impact of using full-text articles (as compared to abstracts) eliminates their advantages. Using full-text will add noise introduced by PDF extraction (and to a lesser extent HTML, too). More important is that the full-text will contain more irrelevant protein and interaction mentions than the abstracts, homonyms of protein names and symbols. These issues will have negatively influenced performance of the full-text based systems. In addition, team 40 used (dependency) parsing, which benefits from clean text data and unbroken sentence structure.

However, even team 40 admitted that of all possible interactions that could be created from normalizing and pairing all proteins mentioned in the abstracts (in total, covering 35% of the training set data), less than two thirds (21% only) of these interaction pairs are found as proteins co-mentioned within one sentence (Rinaldi et al. 2008). Therefore,

[27] to be precise, team 18 in BC II.5 used a clause chunker

their evaluation results indicate that this abstract-based system is already operating at three quarters of the maximum recall possible. They also speculated that if they had invested time into clean full-text extraction from HTML, their approach would probably have performed even better. Furthermore, team 6 measured that in full-text nearly all pairs (95%) can be found co-mentioned in single sentences (Alex, Grover, Haddow, Kabadjov, Klein, Matthews, Tobin & Wang 2008). Team 42 in BC II.5 measured the coverage of pairs within an article section (I.e., creating pairs if the proteins are both present in the section, not just co-mentioned in a sentence.) While 50% of all pairs were recovered from the abstract, 77% were recovered in the results section. They also measured the *proportion* of pair co-mentions in a sentence per section. I.e., they measured where co-mentions are most frequently found in an article. They found that 74% of all pair co-mentions are distributed among three areas, the result (38%) and discussion (17%) sections, plus the figure captions (19%). On the other hand, only 6% of all sentences co-mentioning interaction pairs were attributed to the abstract (Hakenberg et al. 2010)[28].

From the results we can conclude that it is possible to achieve similar performance on abstracts as from full-text, despite the low frequency and coverage of interaction pair co-mentions in abstracts. It should be noted that it is nearly impossible to extract experimental methods from abstracts. Even team 40 used the (HTML) full-text for the method extraction task. In addition, context information used for the disambiguation steps are unlikely to be always present in the abstracts. Equally important, most corpora and available training data for BioNLP systems are composed of ("clean") abstracts text. To text mine entire publications, new corpora for handling these far more noisy texts are being introduced (Bada et al. 2010). While the BioCreative corpora cover thousands of full-text articles, minutely annotated full-texts will require much time and effort to grow to similar sizes. Furthermore, good inter-annotator agreements at these levels of details will be far more difficult to achieve.

5.3.5 Method Extraction Systems

Current approaches reach F-measures of 55%, while using the statistical frequency of the methods to annotate them reaches 36%. However, only BioNLP systems can provide evidence passages from the text to a human annotator.

[28] As an anecdotal note, the heat map they produced from this data made it to the cover of the BioCreative II.5 special issue.

As mentioned in the results of the corpora, a few of the interaction methods appear more frequently than most others. A baseline performance was measured using the most frequent classes in the development set (anti-bait and prey co-immunopreciptation, pull down, and two hybrid). These four most frequent classes, constituting more than 50% of all assignments in the training/development sets, were assigned to each test set article. This statistical baseline has a F-measure of 36% at 29% precision and 48% recall. In other words, only half of the teams were able to substantially improve over this baseline result. However, this baseline assignment is not able to report evidence text passages for its annotations. Therefore, its annotations cannot be interpreted by human curators, and thus have only limited practical value.

Method Keywords as Features

To improve the current performance of text mining systems on this task, the experimental method terms in the PSI-MI ontology need to be associated to the expressions and phrases used in publications.

It has been shown by nearly all participants that the n-grams are the most significant feature in this multi-label classification task. Therefore, the most likely reason why this task is to remain an "unsolved issue", at least if a better performance is expected from the systems, probably has nothing to do with either current text mining methods or technical and computational capabilities.

In ideal circumstances, the set of relevant n-grams - and likely, the related terms, phrases, and patterns, too - that can be used to successfully identify each method could possibly be generated off the spot. If all the tens of thousands of full-text articles curators have meticulously annotated over the years with experimental interaction detection concepts would become available for text processing, these relevant keywords and phrases could be extracted. By grouping the articles according to the curator method annotations and filtering the most distinctive n-grams for each method using an inverted index and a TF-IDF score over the entire set of all annotated articles, the most important features likely would become apparent using one of the most basic statistical BioNLP approaches, at least after a manual inspection of the information-gain- or chi-square-ranked n-grams.

This would not only immediately produce the relevant data needed for this task, but just as much supply the keywords and phrases required to recognize interaction method terms for the protein normalization and interaction detection tasks. It even is thinkable

that full-text article classification then might outperform abstract-based classification. However, there is currently no agreement between the scientific community and the publishing houses that would allow text mining researchers to obtain the necessary collection of articles. Therefore, the only outlook right now is to manually assemble a large collection of phrases that express experimental methods.

5.3.6 Future Outlook

With the current results, the immediate focus should be improving the normalization strategies. That is, enhancing the detection of associations between gene/protein and organism mentions, and possibly even disregarding experimental method relations right now.

With these results at hand, it is obvious that text mining will not be able to reproduce the entire PPI relationships extracted by curators in the foreseeable future. Curators extract not only protein pairs normalized to their DB entries; If present, they annotate entire interaction complexes. Second, they associate the relevant experimental setup with an interaction. Third, they annotate the biological and experimental roles of the interaction participants, a task not part of the BioCreative PPI challenges. This full relationship is the minimal information for PPI curation, as defined by the MIMIx standard.

The foremost goal is the normalization of protein mentions to their correct DB identifiers. And, this target should not be coupled to the requirement that the protein is part of an interaction, much less of an experimental setup, due to the long-range discourse issues. As we have shown, this task is tightly coupled to normalizing the organism mentions. In the latest gene normalization task (BC III), the mention association was not required (i.e., only document-level annotations of gene IDs were measured). The best result for gene normalization had a best break-even score (comparable to the F-measure) of 41%.

However, detecting mentions and making the correct normalizations is tightly coupled. Furthermore, for systems to "learn" the correct normalizations, annotated mentions would provide excellent training data. Therefore, the ideal training data would be semantic annotations on the text, namely the protein/gene mentions and the organism/cell line mentions, associated to their unique identifiers. The evaluation could remain focused on document-level identification of protein/gene and organism mentions, until sufficient performance scores are reached.

A secondary task could be the identification of passages describing experimental evidence. This task's training data would be set up the same way, i.e., passages explicitly annotated with the relevant ontology terms for the experimental method. Furthermore, it would represent the first step in creating the required mappings between the ontology terms and the expressions used in scientific jargon.

Successfully solving these two issues could benefit DB curators and authors annotating publications. It would provide the semantic concept annotations for the publications we read daily. Finally, it will be needed if we wish to fully automate the linking of gene/protein mentions in publications to their database records.

Nonetheless, we wish to discuss some usage scenarios of the current results. Despite initially low results, ways to improve them can be pointed out, as we will outline in the last two sections.

5.4 A PPI Meta-server Framework

The BCMS provides a silver standard for 22,730 abstracts with more than a million (automated) annotations. A meta-service based scenario involving authors, text mining systems, publishers, and reviewers in the DB curation process is outlined.

Here, we will discuss the contributions of the BCMS and its technical feasibility for providing a framework to enable semi-automated author annotations (see Section 5.5). Most of the participant systems are complex pipelines combining individual methodologies, and these systems are developed in-house by various research groups. Therefore, the technical feasibility of a meta-service providing text to and collating annotations from these systems is of interest. In addition, by running the challenges online we hope to motive participants to provide web services to access their often very complex pipelines. (Online, public services have been provided, e.g., by team 42[29] or 10 [30] (BC II.5).)

To address this situation, during BioCreative II a prototype meta-server was developed, able to collect and unify results from distributed systems designed by the BC II research groups. It provides classification and annotations for 22,730 MEDLINE abstracts. The BioCreative MetaServer (BCMS) was an online experiment using web services for collating document annotations. While a complete article annotation system would clearly have to provide more technical functionality, the BCMS is a light demonstration of the viability of such a pipeline.

[29] `http://cbioc.eas.asu.edu/gnat/start.html`
[30] `http://asqa.iis.sinica.edu.tw/biocreative2/`

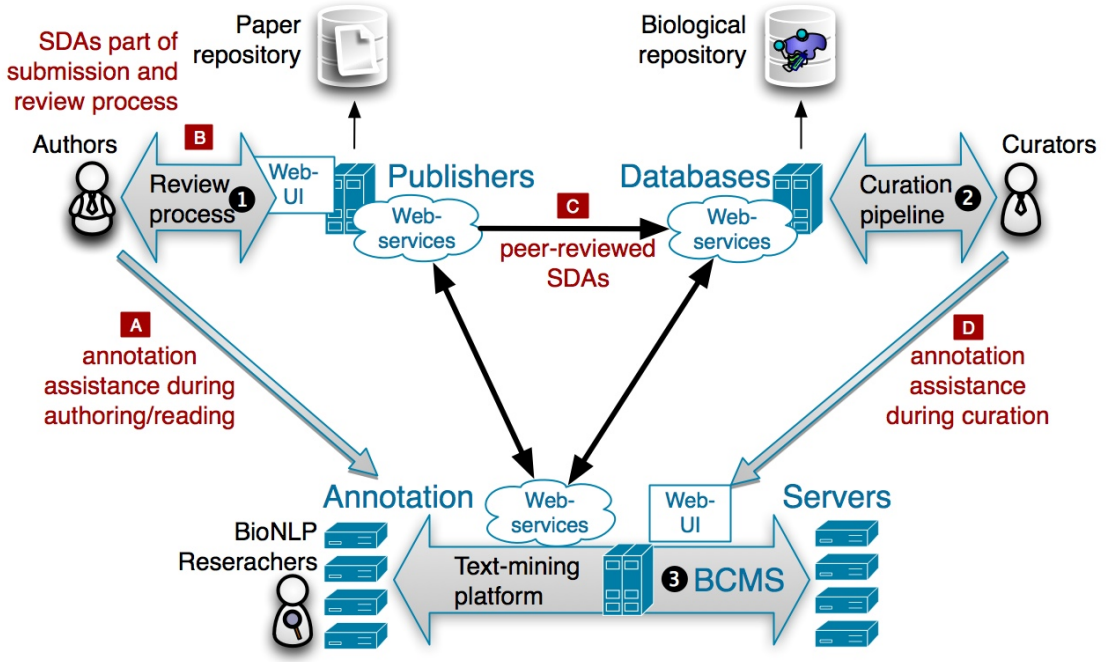

Figure 5.3: A possible scenario for semi-automated author and curator annotations of scientific publications. Authors (1) submit articles to journals (A), that send the articles to a meta-server (3) for annotation. The meta-server distributes the text to BioNLP servers that respond with the annotations. After collating the results and generating a consensus, the publisher site can present the author with potential protein and experimental interaction methods identifiers. The author completes the annotations manually. The submitted, annotated article goes into peer review (B), possibly with the reviewers suggesting improvements to the annotations. After the review process, the articles are published with structured annotations (i.e., SDAs), that can be used by curators (2) to curate the articles, adding the interactions into their repositories (C). Finally, for articles that have not been annotated, curators can request text mining assistance directly, too (D).

The most important, lasting contribution of the BCMS to the scientific community is probably its silver standard annotations. Depending on the desired precision and recall, the over one million automated annotations can be combined with different ensemble approaches, in particular the gene mention data.

An example of a platform that could be used for semi-automated author annotations is shown in Figure 5.3. Both text mining and authors could gain from a "symbiotic" relationship using the author annotations (SDAs) as a feedback to improve or train the results of text mining systems (Leitner & Valencia 2008). However, to scale this experimental prototype to an annotation platform for the publishing industry, it would be necessary to resolve several technical issues, such as supporting all standard web service protocols[31] or adding capabilities to translate between different annotation formats[32]. It would be necessary to add adapters for depositing various document formats[33]. All of these issues are purely of technical nature, while the platform's process design has been presented here (see Section 4.4). As a matter of fact, we experimented with distributed user interfaces on top of the BCMS platform, and the platform's design was ranked third in an annual IT challenge on web services, the ICWS 2009 Services Cup, competing against contenders from industry, such as IBM or SAP (Starlinger et al. 2009).

One possible issue in this respect is the time it takes for automated systems to annotate articles. The best performing system (by team 10) had the slowest run-time, as shown in the BC II.5 challenge' timing results (Figure 4.19). This might be interpreted as an indicator that the time requirement for such a meta-service would be prohibitive. If comparing the normalization approaches of this team with, for example, team 22, it seems unlikely that this system could not be scaled to runtimes that at least equal team 22. That team's pipeline parsed all sentences twice, using a dependency and a (head-driven) phrase-structure grammar (Saetre et al. 2010). Furthermore, given the experimental nature of the online challenge (it was the first entirely online BioNLP community evaluation), the timings of team 10 probably can generally be considered outliers. Furthermore, team 42 experimented with high-throughput approaches and showed that they can annotate articles within a matter of seconds at still acceptable performance (see Figure 4.19).

In other words, based on the challenge data, it should be possible to provide a user with annotations from an online text mining system in about 2 minutes. When an author

[31] SOAP and JSON-REST in addition to the currently supported XML-RPC
[32] e.g., IeXML, UIMA CAS, GENIA XML, GATE XML, or the AO (Annotation Ontology) format
[33] Word, PDF, LaTeX, HTML, OpenDocument Format, RTF, ePub, etc.

first uploads her article to a publisher, she needs to fill in all kinds of forms[34]. By the time she has completed all these forms, a text mining meta-service can have provided the annotations for the article.

5.5 Usability Discussion

This discussion on text mining-assisted curation will focus on the possibilities of implementing a semi-automated annotation strategy for PPI publications.

The traditional separation between database records and journal entries is diminishing (Valencia 2002, Seringhaus & Gerstein 2007). Maintaining consistent annotations in databases, represented by common standards, and linked to related sources and concepts (publications, other databases, and ontologies), are the fundamental requirements to make biological data accessible to other researchers for future analysis (Campbell 2007). These developments could build into a picture in which the publication of papers would be more directly related with the deposition of the relevant information in databases. The published scientific work could be enhanced with links to and from the databases. Two possible scenarios that have been proposed (Hahn et al. 2007) include:

- Allowing authors to freely decide on the annotations, letting them choose concepts from collections of controlled vocabularies (e.g., Gene Ontology (GO) terms). This leaves the obvious difficulty of obtaining consistent annotations from authors not necessarily aware of how to use an ontology. Furthermore, it should be pointed out that training, for example, a GO annotator requires several months of work in close collaboration with experts.

- Semi-automated systems could pre-filter possibly relevant terms from collections of controlled vocabularies and offer the results to human experts for validation. This type of symbiosis could bridge the need between annotation consistency and annotator expertise.

In this discussion, we will expand on the idea of the latter scenario, and show data for the former, as generated by the FEBS Letters experiment. First we will examine how current article classification systems could assist curators in identifying relevant articles, and then assess if protein normalization systems could aid either curators or authors in retrieving the corresponding database identifiers.

[34] author data, affiliations, publication title and abstract, etc.

5.5.1 Article Classification Task

The current article ranking approaches could reduce keyword-based article selection times by at least one third, and can be used to assemble PPI article collections.

The test set class distributions in the BC II.5 and III sets better reflect a real distribution of PPI articles than the balanced situation in BC II. For discussing the usefulness of the results (see Figure 4.4) in a curation scenario, we will focus on the PR curves of the best systems from the two later challenges.

As can be seen from the curves, the best classification systems reach a precision of approximately 70% at 50% recall. The recall is comparable to a keyword-based search: Finding half of the 910 relevant articles by selecting records from the 6000 abstracts in the BC III test set with the best lemma that separates positive from negative articles - "interact" - would retrieve 1215 abstracts, out of which 480 are relevant (i.e., roughly 50% recall, too), while 735 are false positives.

On the other hand, if the articles were read in the order suggested by a classifier, it would require reading the top 11% (650) of all articles to reach 50% recall.[35]. If assuming the keyword-based article selection had the same number of 455 relevant articles (instead of 480), it would contain 697 negative articles[36], for a total of 1152 articles[37]. This means the ranked article set containing only 650 articles nearly halves the number of articles compared to the "interact" keyword-selected article set[38].

However, our measurements of curator annotation times showed that classifying positive articles can require up to twice as much time, depending on curator expertise (Krallinger et al. 2011) – the more experienced the curator, the more equal the classification times are. We can estimate the minimum time saving in the worst case scenario (i.e., taking an upper bound, where positive articles require twice as much curator time) by using the same number of positive articles for both the keyword-based article set and the ranked selection. If assuming the keyword-based article selection has 455 relevant articles, it would contain 697 negative articles (see above). Thus, the relative time curators would need when using a ranked list as compared to a traditional keyword search is 69%[39], or roughly saving them one third of their time.

[35] At an average recall of 50%, half of the 910 true articles in the test set is 455. At 70% precision, that means the system has reported 650 articles to cover these 455 true articles, with 195 false positives.

[36] $735 \cdot 455 \div 480 = 697$

[37] $455 + 697 = 1152$

[38] It has to be noted that keyword-based queries can be improved by query expansion techniques (Efthimiadis 1996), but to our best knowledge there are no comparable evaluation data for query expansion-based PPI article retrieval available.

[39] $(2 \cdot 455 + 195) \div (2 \cdot 455 + 697)$; $2 \cdot 455$ are the doubled time "units" for positive articles, 195 are the

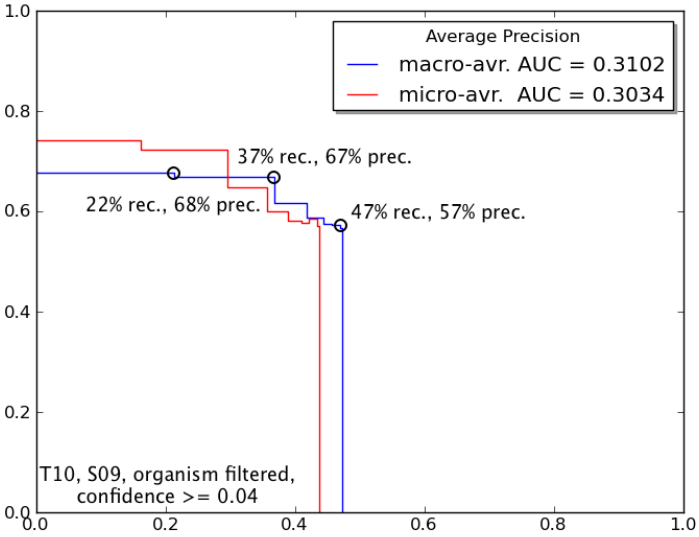

Figure 5.4: AP curve plot of the document-centric, organism filtered protein normalization result of team-run 10-S09 (BC II.5), applying the same cutoff (confidence ≥ 0.04) as in Figure 4.7.

In a similar way, the PR curves can be used to estimate the average number of false positives in a set generated by the classifier. A classifier's ranked list could be used to some predetermined rank or confidence value cutoff to generate a set of PPI articles. For example, team 9 (run S28) in BC II.5 (see Figure 4.4) has a nearly perfect classification for the first 20% of all relevant articles. Thus, if it were enough to extract about 180 PPI-relevant articles from a collection of 6000 that have a similar distribution of relevant articles as the test set, this system could deliver a near perfect set. Put differently, on a set with a proportion of relevant articles as in the test set (15%), the first 3% of the (ranked) articles might all be relevant.

5.5.2 Protein Normalization Task

Organism selection and removal of false positives from a small set of system-provided annotations by a human user could substantially increase the quality of the automated results.

It is hard to argue that the raw results of the protein normalization have a clear usability for semi-automated curation. However, it would be trivial - for curators and

relative time "units" for negative articles in the ranked list, and 697 are the time "units" for the negative articles in the keyword-based list

especially authors - to select the organisms relevant to the document being annotated. In this case, the curator/author would be presented with a normalization result that has the performance properties of the organism filtered, document-centric evaluation. The best normalization systems can produce result sets with a performance of 45-56% F-score (see Figure 4.14), meaning that about half of the reported results will be either correct or present. Furthermore, team 10 (S09) was able to annotate nearly all articles (59/61).

If examining the organism-filtered AP curve of the best system in the two challenges (10-S09, BC II.5, cutoff at confidence 0.04), the author would be confronted with a precision/recall of the ranked results as depicted in Figure 5.4. This curve shows that on average two thirds (67% precision) of the first few results (up to 37% recall) reported by the system are relevant. Approximately half (47% recall) of the (true) proteins in the article are annotated by the system, together with a similar amount of false annotations (57% precision). However, this also means that, on average, the other half of the relevant (true) proteins will not be reported by the system.

If a similar comparison was made with the preliminary ensemble result, these numbers could be even better. An ensemble result could particularly increase the recall, which is important. The more relevant identifiers the automated results can report at reasonable precision, the less manual lookup authors or curators have to make. After organism filtering, an ensemble might achieve 65% document-centric, macro-averaged recall for 89%[40] of all documents. This means an author who has to annotate 5 proteins, only needs to add 2 more to complete the annotation and remove about 2-3 false positives (58% precision).

5.5.3 Comparing Authors, Curators, and Text Mining Systems

Only professional bio-curators reach annotation consistency levels that are sufficient to populate PPI repositories. Nonetheless, incorporating authors and/or text mining systems in the process can increase the quality of those deposits.

The combined FEBS SDA and BioCreative II.5 effort are the first quantitative approach to directly compare the impact of generating comparable annotations for scientific manuscripts using different approaches, namely database curators, manuscript authors, and automated text mining systems (Leitner, Chatr-aryamontri, Mardis, Ceol, Krallinger, Licata, Hirschman, Cesareni & Valencia 2010).

[40] 54 out of 61

The Venn diagrams in Figure 4.16 showed that the results from systems, authors, and curators are disjoint over large parts of their annotations. This makes it likely that the three sources might be able to assist each other (4.16, a). Authors and automated systems have just one quarter of their annotations in common, and joining the results of the best-scoring system with authors' annotations would increase the gold standard coverage of author annotations by 28%/18pp (see Fig. 4.16, b). When joining system and curator annotations (Fig. 4.16, c), the increase in coverage is not as large (5%), but the system reproduces nearly two thirds of the annotations the curators made. Therefore, it might be possible for text mining to speed up the process of retrieving correct protein identifiers for curators, too. In comparison, using the authors' annotations, curators obtain a similar benefit in coverage as the authors' results. And, author annotations have the added advantage of including a much smaller number of false positives when compared to the system annotations (15 vs. 101, Fig. 4.16, d vs. 4.16, c). This last Venn Diagram (4.16, d) visualizes the gain curators had on coverage when basing their work on author data (Figure 4.14, curators vs. authors & curators): The author data, reporting identifiers the curators had missed, allowed the curators to to integrate additional identifiers (missed by the curators when not basing their annotations on author data). However, it should be noted that curators using author annotations did have a slightly lower precision than curators doing their work alone.

It is also worth mentioning that the automated system produced more annotations than the other sources. Some of these additional annotations will refer to proteins mentioned in the article that are not relevant to the task (i.e., have no experimental evidence supporting them). However, these extra results negatively affect calculations of precision and balanced F-measure and are one of the reasons why the text mining systems perform worse than humans (see the precision of the systems in Figure 4.14).

It should be noted that while the system run is for their training set, the system annotations used in the overlap study were not organism filtered. Furthermore, an ensemble result could potentially out-perform a single result, too.

The overlap comparisons showed that the three approaches could assist each other. Especially, systems could help the authors increasing their recall, as here the overlap differences are significantly larger, while authors picking out irrelevant system results can reduce the precision error of systems. Author annotations improved with this semi-automated strategy in turn would benefit curators, resulting in an overall increase of PPI data quality.

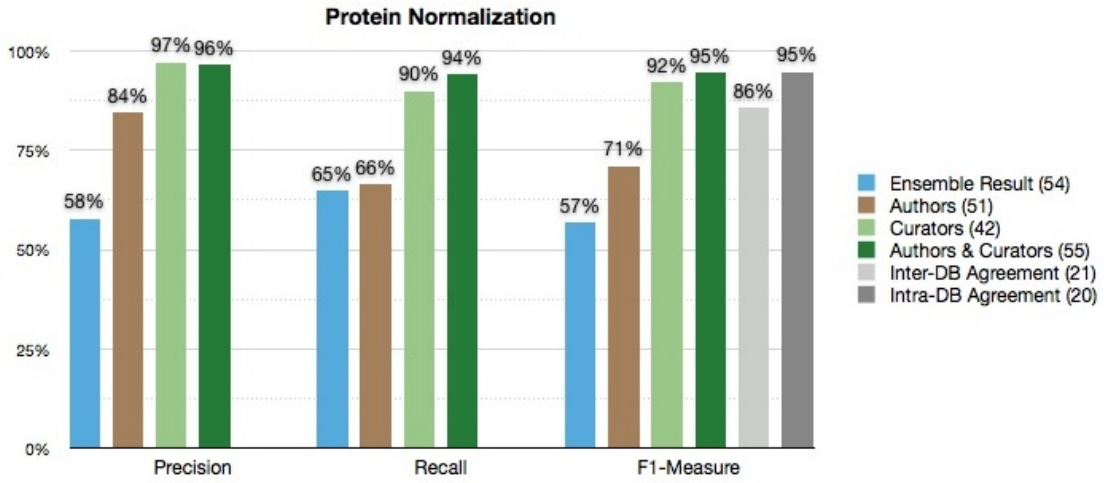

Figure 5.5: Comparison of the potential BioCreative II.5 **ensemble protein normalization** result after document-centric organism filtering (blue) with authors (brown) and curator annotations (green, light: curator alone; green, dark: using author annotations), as well as the IAA (grey, light: inter-DB; grey, dark: intra-DB). Legend numbers in parenthesis: articles annotated.

In Figure 5.5, we replace the individual systems with the ensemble experiment. A suggestive approach therefore would be to make use of a "meta-system", reporting the consensus provided by the best normalization approaches. All authors would need to do to produce equally competent annotations of their own would be:

- Select the relevant organisms for the interacting proteins presented in their article.
- Remove a few of the remaining false positive predictions made by the systems.

A beneficial side-effect of a semi-automated annotation process is that it might direct authors towards using the official gene names and symbols. An author using the correct names will be rewarded with better automated annotations, and the benefit of using less ambiguous names will be noticed by readers, too. Second, if authors used gene names more uniformly, search engine keyword queries for a specific gene or protein would be far more likely to retrieve all relevant publications. Even if authors do not wish to change the names used in their publications, the annotation system would inform the author or publisher about the official gene and protein symbols to declare as keywords on the article.

5.5.4 Annotation of PPI Articles

Out of 76 authors participating in the FEBS Letters experiment on SDAs, three quarters seemed to be willing to contribute to the PPI annotation effort. And, a semi-automated approach is likely to increase the coverage of author annotations.

Manual bio-curation is a slow process and databases lag behind, not able cover the entire space of published PPI information. One way to extend the number of interactions captured by public repositories is to enlist authors in the annotation effort. As suggested by (Orchard et al. 2007), authors could be asked to submit the relevant information during the editorial process, as defined by the minimum information requirement for reporting protein interaction experiments (MIMIx). The pros and cons of possible approaches for adding structured information to scientific publications have been discussed: (Gerstein et al. 2007) and (Hahn et al. 2007) argued over the extent to which quality, consistency, and stable support could be expected from authors, and to what extent automatic systems can help in this process. However, the debate had been stalling because of the absence of comparable data about these approaches.

To complement the FEBS Letters SDA experiment on author annotations, the BioCreative organizers challenged text-mining researchers to reproduce a subset of these PPI annotations. From the analysis of the combined results we can draw the following conclusions on five critical issues (Leitner, Chatr-aryamontri, Mardis, Ceol, Krallinger, Licata, Hirschman, Cesareni & Valencia 2010):

1. Including authors in the curation process. The majority of authors publishing PPI articles in FEBS Letters during the period of the experiment (76 in total) were willing and able to provide structured information: 57 of the authors agreed to participate in the experiment, 56 provided curation-relevant PPI data, and 17 out of 22 authors who responded to a questionnaire accompanying the experiment expressed their interest in SDAs (see Section 4.3.7).

2. Increasing the quality of curated data. Possibly due to the comparatively low accuracy of authors' submissions, the use of author annotations did not result in a saving of curator time, but the data quality increases relative to their individual results (see Figures 4.14 and 4.15, "curator" vs. "authors & curator"). Particularly, the coverage increased by 4 and 5pp (normalizations and interactions, respectively). This claim is also consolidated by the investigation of the overlap of annotations between the three sources (see the Results in Section 4.3.6).

3. Interest in and technical feasibility of providing SDAs by journals. Adding SDAs to the editorial process was not only desirable for the FEBS editorial house, but also technically possible. These SDAs continue to form part of the two FEBS, i.e., FEBS Letters (Elsevier) and FEBS Journal (Blackwell). Furthermore, journals incorporating SDAs are likely to increase the impact and visibility of their articles: unambiguously labeled articles will be more relevant in search results.

4. Using automated systems to assist in the annotation process. The (online) systems required on average 2.3 minutes (std. dev. = 1.4 min) to provide annotations on the test set articles, while 22 of the authors reported spending 68 minutes (std. dev. = 72 min) on average. Measurements of MINT curators over 36 PPI articles show an average of 50 minutes (std. dev. = 26 min) per article[41]. According to both authors and curators, retrieving the correct identifiers is the most time-consuming task. Therefore, having automated systems provide a list of relevant identifiers is *likely* to reduce annotation time. As shown in the overlap results, systems are able to return identifiers that both authors and curators miss, indicating that using a semi-automated approach is *likely* to increase coverage. And, combining the identifiers reported by multiple automated systems into a *preliminary* ensemble result achieved a similar coverage to the authors'.

5. In addition, from the BioCreative article classification results (Section 4.3.1), we established that text mining can facilitate identifying PPI-reporting articles for database curators. As shown in this work, the text mining systems could reduce the time spent by curators on retrieving relevant articles between a third (worst estimate) and a half (best estimate).

However, given the results of the systems on interaction pairs, it is also clear that complex curation tasks cannot be transferred to automated approaches. And, the relatively small samples of articles in this study has to be acknowledged as well. Nonetheless, with our evaluation results it seems possible that integrating text mining in the process can provide authors with protein identifiers requiring very little human intervention in the future, and it might be possible that this would save authors' time, too. From the overlap study we would expect that at the very least the systems would help increase the coverage of author annotations, because the system was able to annotate 17%[42] of

[41] Established by data provided to us from Gianni Cesareni taking measurements of MINT curation times.

[42] $27 \div (27 + 60 + 35 + 38)$, see Figure 4.16, b

the gold standard identifiers that were not provided by the authors.

Therefore, it would be interesting to compare annotations from curators or authors using different approaches in future interactive tasks of BioCreative. As already indicated, the critical question will be how much time an annotator has to spend discarding the false positive results and how difficult the remaining false negatives are to annotate. These measurements will be needed to predict the actual time saving capabilities. It cannot be ruled out that such a system could, in the worst case, increase annotation time or decrease annotation quality. However, another study with PPI curators concluded that (text mining-) assisted curation did decrease curation time (Alex, Grover, Haddow, Kabadjov, Klein, Matthews, Roebuck, Tobin & Wang 2008). Furthermore, the curator annotations based on author data did show increased recall (see green bars in Figures 4.14 and 4.15).

In the next BioCreative interactive tasks it therefore would be interesting to directly confront curators and authors with the normalization systems. The goal would be to compare the average time spent annotating an article with and without the help of a text mining or an ensemble approach, as well as compare the average quality of the annotations between the settings.

CHAPTER 6

Conclusions

6.1 English

This thesis contains the description of our assessment of current text mining approaches and an estimation of their possible benefit for manual annotation of protein interactions by database curators and authors. The results presented correspond to the three BioCreative community challenges we have co-organized. In particular, they are based on BioCreative tasks related to protein interaction extraction and retrieval.

The main conclusions of this work are:

1. We have evaluated 314 submissions from 52 research groups with respect to article classification, protein normalization, interaction detection, and experimental method extraction.

2. Several PPI gold corpora, assembled for the challenges, have become standard in the field. They form a collection of 19,642 abstracts and 3,632 full-text articles freely available to the scientific community to continue developing IR/IE approaches related to PPIs.

3. The BioCreative Meta-Server (BCMS) has been shown to be an effective approach to collate information from distributed servers. The current results for 22,730 MEDLINE abstracts - annotated by 13 servers - provide a silver standard with annotations for proteins, genes, and taxa.

4. We have shown that automated article ranking can improve PPI-relevant article selection. Automated ranking can reduce the time curators would need by at least one third if compared to a traditional keyword-based retrieval without PPI-specific ranking.

5. The results show that organism disambiguation is the most prevalent difficulty in protein mapping approaches. Mouse, fly, and Arabidopsis proteins are particularly hard to identify correctly, while rat proteins are not. Selecting the relevant focus organisms could double the raw performance of protein normalization results.

6. In a relevant number of cases, PPIs are extracted (by curators) from long-range discourse relationships (between proteins, organisms, and/or experiments) and their detection is beyond the capabilities of current natural language processing methods.

7. A new score was introduced. The FAP-score, composed as the harmonic mean of F-measure and Average Precision (AP), was clearly able to identify the best results of the BioCreative challenges. Compared to previously existing scoring metrics, it does not require setting an artificial threshold and it was shown to be better at penalizing systems generating lists with many irrelevant results.

8. The combination of results from the best methods, together with the new FAP-score and the use of meta-services, opens new avenues for a more effective text mining, able to better assist authors and curators in retrieving the relevant identifiers for the interacting proteins.

6.2 Castellano (Spanish)

Esta tesis contiene la descripción de nuestra evaluación de los enfoques actuales de minería de textos, y su posible beneficio para el proceso de anotación manual de interacciones de proteínas llevado a cabo por curadores y autores. Los resultados presentados se atribuyen a los tres competiciones de BioCreative que hemos coorganizado. Concretamente, están basados en las tareas de BioCreative relacionadas con la extracción y recuperación de información (IE/IR) sobre interacciones entre proteínas (PPIs).

Las conclusiones principales de este trabajo son:

1. Hemos evaluado 314 resultados generados por 52 grupos de investigación para las tareas de clasificación de artículos, la normalización de las proteínas, detección de interacciones, y a la extracción de métodos experimentales.

2. Varios corpora "gold standard" de PPI, producidos para las competiciones, se han convertido en un estándar en el campo. Están constituidos por una colección de 19.642 resúmenes y de 3.632 artículos con texto completo, los cuales están accesibles libremente para la comunidad científica con el objetivo de garantizar la continuación del desarrollo de sistemas de IE/IR relacionados con PPIs.

3. Se ha podido mostrar que el BioCreative Meta-Server (BCMS) constituye una estrategia eficaz para unir la información generada por servidores distribuidos. Los resultados actuales para 22.730 resúmenes de MEDLINE, los cuales han sido anotados por 13 servidores, proporcionan un "silver standard" con anotaciones de proteínas, genes y especies.

4. Hemos mostrado que el "ranking" automático de artículos de PPI puede resultar en una mejora en la selección de artículos. La clasificación automática es capaz de reducir el tiempo que los curadores necesitan para este proceso por lo menos en un tercio si se compara con una recuperación tradicional basada en una búsqueda con palabras clave.

5. Los resultados indican que la desambiguación de organismos es la dificultad más frecuente en la asignación de identificadores a menciones de proteínas. Las proteínas de ratón, mosca y Arabidopsis son particularmente difíciles de identificar correctamente, mientras que las de las ratas no lo son. Una selección previa de los organismos es capaz de duplicar el rendimiento básico de los resultados generados en el caso de la normalización de proteínas.

6. En un número relevante de casos, las PPIs se encuentran descritas a través de relaciones de discurso a distancia (entre proteínas, organismos, y/o experimentos), y su detección está fuera del alcance de los métodos actuales de procesamiento del lenguaje natural.

7. Hemos introducido una nueva métrica de evaluación. El valor FAP, el cual consiste en la media armónica del "F-score" y el "Average Precision" (AP), fue claramente capaz de detectar los mejores resultados en los competiciones BioCreative. En comparación con las métricas de evaluación previamente utilizadas, no requiere de un valor de corte artificial y ha demostrado ser mejor a la hora de penalizar sistemas que generan listas que contienen muchos resultados irrelevantes.

8. La combinación de los resultados de los mejores métodos, junto con el nuevo valor FAP y el uso de meta-servicios, abre nuevos caminos para una minería de textos más eficaz, capaz de asistir mejor a los autores y curadores en la tarea de recuperar identificadores de proteínas que participan en interacciones.

APPENDIX A

Appendix

A.1 Detailed Evaluation Results

A.1.1 Article Classification Results

BC	Team	Run	sens.	spec.	acc.	MCC	AUC PR
II	04	1	0.83728	0.64897	0.74298	0.49503	0.68149
II	04	2	0.93787	0.39233	0.66470	0.39384	0.48381
II	04	3	0.79290	0.68142	0.73708	0.47725	0.69700
II	06	1	0.86095	0.64602	0.75332	0.51901	0.85423
II	07	1	0.84320	0.61357	0.72821	0.46923	0.81573
II	07	2	0.86391	0.57227	0.71787	0.45590	0.75113
II	07	3	0.85799	0.60472	0.73117	0.47821	0.82453
II	11	1	0.86686	0.51622	0.69129	0.40897	0.79180
II	11	2	0.76923	0.70501	0.73708	0.47520	0.78229
II	11	3	0.78107	0.62832	0.70458	0.41421	0.63317
II	14	1	0.44970	0.83776	0.64402	0.31196	0.77192
II	14	2	0.46450	0.83481	0.64993	0.32227	0.77629
II	14	3	0.47041	0.84071	0.65583	0.33500	0.77917
II	19	1	0.73373	0.56047	0.64697	0.29869	0.64470
II	19	2	0.56509	0.69027	0.62777	0.25739	0.64470
II	27	1	0.85503	0.40413	0.62925	0.29029	0.62040
II	27	2	0.92012	0.26549	0.59232	0.24545	0.57312
II	27	3	0.85207	0.45133	0.65140	0.33108	0.63795
II	28	1	0.81065	0.73156	0.77105	0.54388	0.83536
II	28	2	0.76923	0.74041	0.75480	0.50984	0.81013
II	28	3	0.78994	0.64012	0.71492	0.43493	0.80012
II	30	1	0.59467	0.57522	0.58493	0.16993	0.60569
II	30	2	0.50296	0.47198	0.48744	-0.02508	0.57416
II	30	3	0.69527	0.53687	0.61595	0.23509	0.65439
II	31	1	0.59467	0.70501	0.64993	0.30155	0.68009
II	31	2	0.52663	0.79646	0.66174	0.33558	0.69951
II	31	3	0.34615	0.91150	0.62925	0.31247	0.68849
II	37	1	0.97929	0.19469	0.58641	0.28046	0.65028
II	37	2	0.94675	0.30383	0.62482	0.32702	0.70199
II	37	3	0.98225	0.13569	0.55835	0.22145	0.63132
II	41	1	0.88757	0.43363	0.66027	0.36039	0.71583
II	41	2	0.87574	0.45428	0.66470	0.36383	0.75412
II	41	3	0.89053	0.45428	0.67208	0.38309	0.74298
II	44	1	0.85799	0.61357	0.73560	0.48622	0.69657
II	44	2	0.85799	0.53097	0.69424	0.41150	0.52241
II	44	3	0.82544	0.66077	0.74298	0.49288	0.70566
II	48	1	0.09172	0.99115	0.54210	0.18971	0.65935
II	48	2	0.86391	0.39823	0.63072	0.29614	0.58003
II	48	3	0.31361	0.93805	0.62629	0.32232	0.71342

BC	Team	Run	sens.	spec.	acc.	MCC	AUC PR
II	49	1	0.98521	0.11504	0.54948	0.20334	0.80146
II	49	2	0.99112	0.07670	0.53323	0.16747	0.80022
II	49	3	0.72189	0.46608	0.59380	0.19443	0.58142
II	52	1	0.83432	0.63127	0.73264	0.47542	0.79461
II	52	2	0.84615	0.57522	0.71049	0.43767	0.81289
II	57	1	0.87574	0.63127	0.75332	0.52277	0.79638
II	57	2	0.87278	0.63127	0.75185	0.51933	0.79364
II	57	3	0.86095	0.62537	0.74298	0.50031	0.78301
II	58	1	0.43491	0.86726	0.65140	0.33519	0.72074
II	58	2	0.50296	0.84956	0.67651	0.37589	0.74584
II	58	3	0.73077	0.63717	0.68390	0.36954	0.73681
II.5	07	S04	0.95238	0.48684	0.53613	0.27223	0.44918
II.5	09	1	0.66667	0.79887	0.78487	0.33061	0.26384
II.5	09	2	0.69841	0.77820	0.76975	0.32945	0.27288
II.5	09	3	0.66667	0.78571	0.77311	0.31647	0.25128
II.5	09	4	0.66667	0.86278	0.84202	0.41255	0.53926
II.5	09	5	0.66667	0.85150	0.83193	0.39611	0.55249
II.5	09	S14	0.52381	0.96241	0.91597	0.52519	0.63980
II.5	09	S26	0.69841	0.90789	0.88571	0.51372	0.60769
II.5	09	S27	0.31746	0.99060	0.91933	0.47245	0.55829
II.5	09	S28	0.41270	0.98120	0.92101	0.50834	0.67165
II.5	09	S29	0.69841	0.93797	0.91261	0.58336	0.66779
II.5	13	1	1.00000	0.00000	0.10588	0.00000	0.54841
II.5	13	2	1.00000	0.00000	0.10588	0.00000	0.41904
II.5	13	3	1.00000	0.00000	0.10588	0.00000	0.54983
II.5	13	4	1.00000	0.00000	0.10588	0.00000	0.56941
II.5	13	5	1.00000	0.00000	0.10588	0.00000	0.61211
II.5	14	1	0.90476	0.57331	0.60840	0.29450	0.19247
II.5	14	2	0.73016	0.65789	0.66555	0.24559	0.19459
II.5	14	3	0.95238	0.38534	0.44538	0.21792	0.14614
II.5	14	4	0.74603	0.63722	0.64874	0.24037	0.19363
II.5	14	5	0.96825	0.02632	0.12605	-0.01033	0.10646
II.5	14	S08	0.88889	0.34211	0.40000	0.15266	0.19126
II.5	14	S25	0.88889	0.34087	0.39899	0.15210	0.19126
II.5	16	1	0.34921	0.93045	0.86891	0.28790	0.35828
II.5	16	2	0.26984	0.90789	0.84034	0.17414	0.26341
II.5	16	3	0.33333	0.92857	0.86555	0.26963	0.37548
II.5	16	4	0.52381	0.94361	0.89916	0.46742	0.52973
II.5	16	5	0.42857	0.93233	0.87899	0.36090	0.44139
II.5	20	S32	0.58730	0.95677	0.91765	0.55594	0.68052
II.5	20	S33	0.85714	0.84211	0.84370	0.50975	0.64928
II.5	31	1	0.53968	0.95865	0.91429	0.52511	0.37118
II.5	31	2	0.53968	0.95865	0.91429	0.52511	0.48824
II.5	31	3	0.20635	0.99624	0.91261	0.39763	0.30262

BC	Team	Run	sens.	spec.	acc.	MCC	AUC PR
II.5	31	S18	0.57143	0.94361	0.90420	0.50462	0.40128
II.5	32	S07	1.00000	0.00000	0.24590	0.00000	0.12686
III	65	1	0.38571	0.97642	0.88683	0.48297	0.63463
III	65	2	0.59231	0.93065	0.87933	0.52727	0.63506
III	65	3	0.83077	0.64185	0.67050	0.34244	0.41414
III	65	4	0.71209	0.74126	0.73683	0.34650	0.41414
III	65	5	0.52198	0.94401	0.88000	0.50255	0.61893
III	70	1	0.94176	0.49695	0.56445	0.31789	0.56462
III	70	2	0.38791	0.96108	0.87410	0.43346	0.56462
III	70	3	0.72527	0.83605	0.81924	0.46563	0.56462
III	70	4	0.96593	0.39041	0.47774	0.27060	0.56462
III	70	5	0.20989	0.98624	0.86843	0.34488	0.56462
III	73	1	0.63736	0.91807	0.87550	0.53524	0.65772
III	73	2	0.56703	0.94951	0.89150	0.55306	0.67807
III	73	3	0.60769	0.92613	0.87783	0.52932	0.65736
III	73	4	0.58352	0.94342	0.88883	0.55054	0.67830
III	73	5	0.62088	0.92181	0.87617	0.53031	0.65166
III	81	1	0.60549	0.58762	0.59033	0.13949	0.18645
III	81	2	0.61868	0.57859	0.58467	0.14219	0.18378
III	81	3	0.84945	0.14715	0.25367	-0.00344	0.15102
III	81	4	0.31538	0.69155	0.63450	0.00538	0.15667
III	81	5	0.23407	0.77348	0.69167	0.00645	0.14961
III	81	S09	0.00440	0.99980	0.84883	0.05220	0.43514
III	81	S10	0.05824	0.99607	0.85383	0.17771	0.49335
III	81	S11	0.00110	0.99862	0.84733	-0.00272	0.44650
III	81	S12	0.02857	0.98861	0.84300	0.05244	0.30643
III	81	S13	0.00769	0.99921	0.84883	0.05791	0.17446
III	88	1	0.84725	0.35108	0.42633	0.15238	0.21535
III	88	2	0.74725	0.53733	0.56917	0.20417	0.25433
III	89	1	0.75055	0.80904	0.80017	0.44911	0.61058
III	89	2	0.76813	0.81749	0.81000	0.47242	0.61773
III	89	3	0.74286	0.83851	0.82400	0.48180	0.60112
III	89	4	0.48242	0.94794	0.87733	0.47967	0.42906
III	89	5	0.61868	0.91807	0.87267	0.52082	0.47542
III	89	S04	0.77582	0.77839	0.77800	0.43152	0.57255
III	89	S05	0.77473	0.78153	0.78050	0.43424	0.57318
III	89	S06	0.73736	0.81002	0.79900	0.44073	0.54627
III	89	S07	0.53736	0.92063	0.86250	0.46156	0.40059
III	89	S08	0.67143	0.90393	0.86867	0.53336	0.45546
III	90	1	0.52857	0.95147	0.88733	0.52736	0.50349
III	90	2	0.53626	0.94971	0.88700	0.52890	0.50851
III	90	3	0.56923	0.93929	0.88317	0.52914	0.65065
III	90	4	0.49231	0.96031	0.88933	0.52237	0.48221
III	90	5	0.52527	0.95049	0.88600	0.52204	0.49994

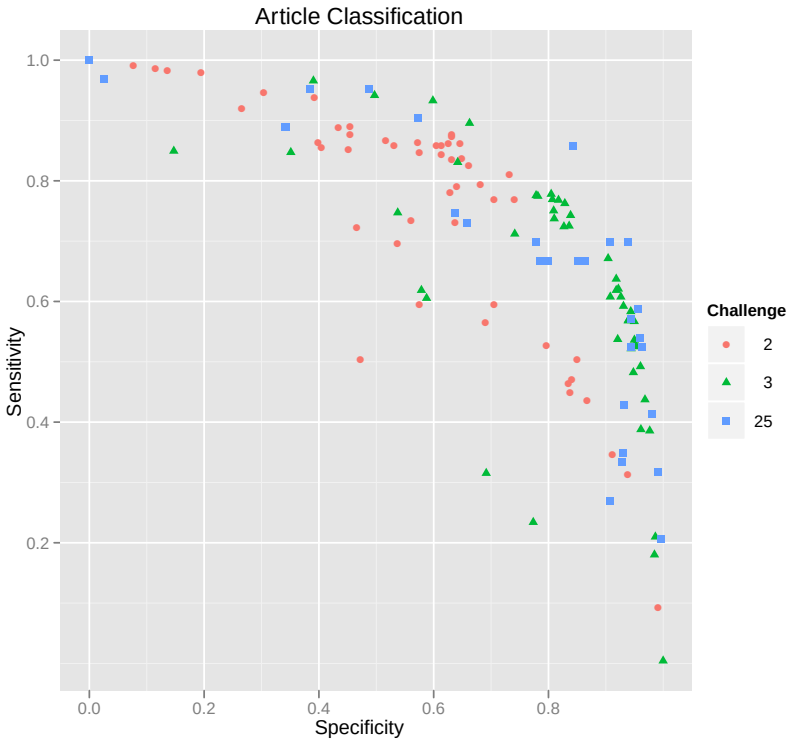

Figure A.1: Plot of sensitivity vs. specificity for all article classification results. Red, circles: BioCreative II submissions; green, triangles: BioCreative III submissions; blue, boxes: BioCreative II.5 submissions.

BC	Team	Run	sens.	spec.	acc.	MCC	AUC PR
III	92	1	0.60769	0.90766	0.86217	0.49155	0.50513
III	100	1	0.43736	0.96817	0.88767	0.50005	0.61468
III	100	2	0.56813	0.93890	0.88267	0.52732	0.61690
III	100	3	0.72418	0.82692	0.81133	0.45256	0.60109
III	100	4	0.76264	0.82849	0.81850	0.48270	0.63619
III	104	1	0.76923	0.80688	0.80117	0.45999	0.53354
III	104	2	0.77802	0.80472	0.80067	0.46370	0.53354
III	104	3	0.93297	0.59862	0.64933	0.38161	0.53354
III	104	4	0.89560	0.66248	0.69783	0.40530	0.53354
III	104	5	0.18022	0.98468	0.86267	0.30064	0.53354

A.1.2 Macro-averaged Protein Normalization Results

BC	Team	Run	prec.	rec.	F_1	AP	FAP
II	4	1	0.1483	0.1126	0.1201	n/a	n/a
II	6	1	0.2447	0.2516	0.2307	n/a	n/a
II	6	2	0.2393	0.2503	0.2232	n/a	n/a
II	6	3	0.2584	0.2979	0.2498	n/a	n/a
II	11	1	0.1219	0.3749	0.1646	n/a	n/a
II	11	2	0.1219	0.3749	0.1646	n/a	n/a
II	11	3	0.1233	0.3739	0.1654	n/a	n/a
II	14	1	0.2209	0.1927	0.1895	n/a	n/a
II	14	2	0.1842	0.1418	0.1502	n/a	n/a
II	14	3	0.1794	0.1752	0.1640	n/a	n/a
II	17	1	0.0837	0.4066	0.1306	n/a	n/a
II	17	2	0.1735	0.3078	0.2058	n/a	n/a
II	17	3	0.2179	0.2657	0.2221	n/a	n/a
II	19	1	0.1808	0.3180	0.2113	n/a	n/a
II	19	2	0.2024	0.3631	0.2384	n/a	n/a
II	19	3	0.2141	0.2863	0.2258	n/a	n/a
II	28	1	0.2215	0.3369	0.2454	n/a	n/a
II	28	2	0.2183	0.2426	0.2122	n/a	n/a
II	28	3	0.2442	0.2150	0.2142	n/a	n/a
II	30	1	0.1119	0.3807	0.1594	n/a	n/a
II	30	2	0.0761	0.4569	0.1234	n/a	n/a
II	30	3	0.2145	0.2591	0.2103	n/a	n/a
II	36	1	0.1220	0.2241	0.1462	n/a	n/a
II	36	2	0.0855	0.2374	0.1134	n/a	n/a
II	36	3	0.1290	0.1941	0.1406	n/a	n/a
II	40	1	0.2180	0.3503	0.2471	n/a	n/a
II	40	2	0.3298	0.2902	0.2892	n/a	n/a
II	42	1	0.0645	0.5492	0.1085	n/a	n/a
II	42	2	0.2501	0.2196	0.2185	n/a	n/a
II	42	3	0.2456	0.2239	0.2190	n/a	n/a
II	42	4	0.1213	0.4935	0.1785	n/a	n/a
II	43	1	0.1072	0.2244	0.1306	n/a	n/a
II	43	2	0.1505	0.1290	0.1294	n/a	n/a
II	43	3	0.1456	0.1745	0.1454	n/a	n/a
II	47	1	0.1845	0.2801	0.2016	n/a	n/a
II	47	2	0.1884	0.2758	0.2027	n/a	n/a
II	47	3	0.1831	0.2898	0.2021	n/a	n/a
II	49	1	0.0594	0.2658	0.0913	n/a	n/a
II	49	2	0.1050	0.1594	0.1176	n/a	n/a
II	49	3	0.0841	0.1840	0.1056	n/a	n/a
II	58	1	0.0271	0.0757	0.0364	n/a	n/a

BC	Team	Run	prec.	rec.	F_1	AP	FAP
II	58	2	0.0266	0.0744	0.0353	n/a	n/a
II	58	3	0.0354	0.0717	0.0422	n/a	n/a
II	60	1	0.0923	0.2325	0.1189	n/a	n/a
II	60	2	0.0365	0.0738	0.0450	n/a	n/a
II	60	3	0.0543	0.0874	0.0626	n/a	n/a
II.5	10	1	0.01245	0.65193	0.02380	0.26607	0.04369
II.5	10	2	0.01248	0.65193	0.02386	0.26423	0.04376
II.5	10	3	0.01239	0.64647	0.02368	0.26419	0.04347
II.5	10	4	0.01245	0.65193	0.02380	0.26432	0.04367
II.5	10	5	0.01264	0.67632	0.02418	0.26283	0.04428
II.5	10	S09	0.08748	0.59101	0.14508	0.26213	0.18678
II.5	10	S21	0.01457	0.64647	0.02788	0.26575	0.05047
II.5	10	S23	0.01469	0.64647	0.02811	0.26019	0.05074
II.5	10	S24	0.01457	0.64647	0.02788	0.26501	0.05045
II.5	14	1	0.27869	0.26739	0.25457	0.09804	0.14157
II.5	14	2	0.20453	0.20863	0.19414	0.06739	0.10005
II.5	14	3	0.18858	0.35331	0.22052	0.12533	0.15982
II.5	14	4	0.18913	0.21919	0.19557	0.06875	0.10174
II.5	14	S08	0.12951	0.12704	0.11725	0.02673	0.04354
II.5	14	S25	0.17780	0.17575	0.16048	0.04291	0.06772
II.5	18	5	0.22216	0.40130	0.26673	0.13954	0.18322
II.5	22	1	0.02910	0.59566	0.05452	0.21351	0.08686
II.5	22	2	0.00613	0.67895	0.01208	0.20233	0.02279
II.5	22	3	0.02906	0.59566	0.05446	0.22615	0.08779
II.5	22	4	0.00646	0.68305	0.01271	0.22216	0.02405
II.5	22	5	0.02019	0.54392	0.03842	0.16821	0.06256
II.5	22	S05	0.02779	0.58396	0.05239	0.16008	0.07894
II.5	22	S10	0.01915	0.58311	0.03641	0.09665	0.05289
II.5	22	S11	0.02781	0.57647	0.05238	0.16521	0.07954
II.5	22	S12	0.03115	0.57920	0.05753	0.15011	0.08318
II.5	22	S13	0.02175	0.51308	0.04119	0.12160	0.06153
II.5	26	S16	0.02279	0.14861	0.03714	0.01117	0.01718
II.5	31	1	0.07541	0.20826	0.10567	0.01815	0.03097
II.5	31	2	0.07541	0.20826	0.10567	0.01672	0.02887
II.5	31	3	0.00899	0.58160	0.01760	0.00508	0.00789
II.5	31	4	0.01460	0.26814	0.02705	0.00382	0.00670
II.5	31	5	0.07869	0.21304	0.10981	0.01798	0.03089
II.5	31	S18	0.09016	0.22119	0.12171	0.02133	0.03631
II.5	32	1	0.06817	0.44374	0.11296	0.05643	0.07526
II.5	32	S07	0.06658	0.41997	0.10992	0.05001	0.06874
II.5	37	1	0.18089	0.41319	0.20009	0.09292	0.12690
II.5	37	2	0.02535	0.60360	0.04778	0.01604	0.02402
II.5	37	3	0.19587	0.44516	0.22783	0.12444	0.16096
II.5	37	4	0.03084	0.63571	0.05768	0.02060	0.03036

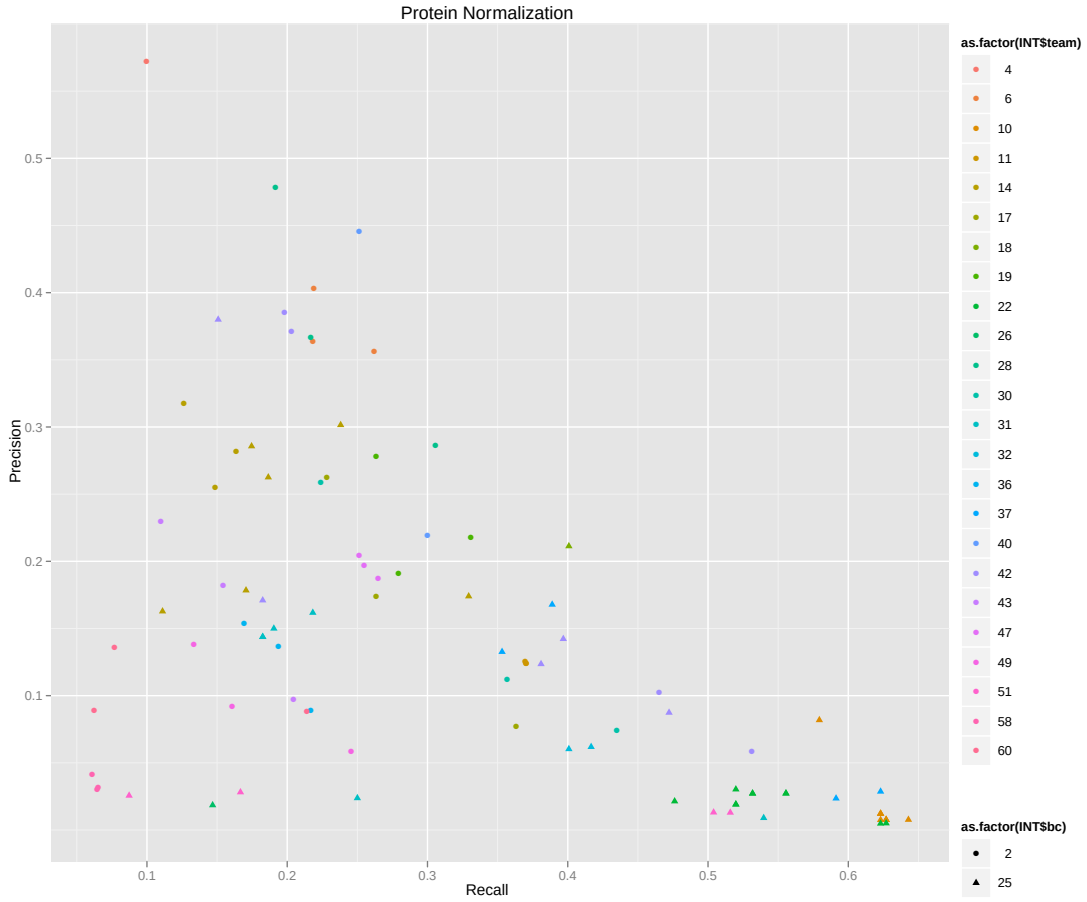

Figure A.2: Micro-averaged precision vs. recall plots of all protein normalization results. Triangles: BioCreative II.5 submissions, circles: BC II. Runs are color-coded per team.

BC	Team	Run	prec.	rec.	F_1	AP	FAP
II.5	42	S01	0.14945	0.16608	0.14771	0.02901	0.04849
II.5	42	S02	0.15326	0.43597	0.20098	0.13378	0.16064
II.5	42	S03	0.13713	0.42867	0.18983	0.11732	0.14501
II.5	42	S19	0.12470	0.21293	0.14433	0.04087	0.06370
II.5	42	S20	0.09486	0.48244	0.15021	0.13481	0.14210
II.5	51	1	0.00634	0.09133	0.01179	0.00502	0.00704
II.5	51	2	0.01252	0.17695	0.02318	0.01363	0.01717
II.5	51	3	0.01856	0.51938	0.03535	0.03766	0.03647
II.5	51	4	0.01806	0.52995	0.03446	0.03775	0.03603

A.1.3 Macro-averaged Interaction Detection Results

BC	Team	Run	prec.	rec.	F_1	AP	FAP
II	4	1	0.0988	0.0672	0.0664	n/a	n/a
II	6	1	0.1493	0.1514	0.1322	n/a	n/a
II	6	2	0.1256	0.1367	0.1124	n/a	n/a
II	6	3	0.1380	0.1702	0.1298	n/a	n/a
II	11	1	0.0353	0.1748	0.0484	n/a	n/a
II	11	2	0.0353	0.1748	0.0484	n/a	n/a
II	11	3	0.0355	0.1748	0.0487	n/a	n/a
II	14	1	0.1106	0.0851	0.0837	n/a	n/a
II	14	2	0.0939	0.0622	0.0677	n/a	n/a
II	14	3	0.0701	0.0611	0.0578	n/a	n/a
II	17	1	0.0330	0.1928	0.0497	n/a	n/a
II	17	2	0.0774	0.1468	0.0879	n/a	n/a
II	17	3	0.1073	0.1311	0.1029	n/a	n/a
II	19	1	0.0631	0.1532	0.0758	n/a	n/a
II	19	2	0.0829	0.1907	0.0975	n/a	n/a
II	19	3	0.1078	0.1587	0.1109	n/a	n/a
II	28	1	0.0908	0.1775	0.1012	n/a	n/a
II	28	2	0.1144	0.1264	0.1025	n/a	n/a
II	28	3	0.1383	0.1188	0.1119	n/a	n/a
II	30	1	0.0459	0.1491	0.0596	n/a	n/a
II	30	2	0.0282	0.1794	0.0422	n/a	n/a
II	30	3	0.1011	0.1116	0.0850	n/a	n/a
II	36	1	0.0295	0.0761	0.0353	n/a	n/a
II	36	2	0.0167	0.0659	0.0209	n/a	n/a
II	36	3	0.0324	0.0617	0.0368	n/a	n/a
II	40	1	0.0582	0.1757	0.0740	n/a	n/a
II	40	2	0.1648	0.1494	0.1329	n/a	n/a
II	42	1	0.0130	0.3055	0.0226	n/a	n/a
II	42	2	0.1304	0.1104	0.1034	n/a	n/a
II	42	3	0.1212	0.1055	0.0996	n/a	n/a
II	42	4	0.0310	0.2553	0.0465	n/a	n/a
II	43	1	0.0294	0.0668	0.0325	n/a	n/a
II	43	2	0.0458	0.0368	0.0355	n/a	n/a
II	43	3	0.0407	0.0560	0.0381	n/a	n/a
II	47	1	0.0590	0.1263	0.0624	n/a	n/a
II	47	2	0.0618	0.1245	0.0636	n/a	n/a
II	47	3	0.0547	0.1271	0.0595	n/a	n/a
II	49	1	0.0091	0.0807	0.0151	n/a	n/a
II	49	2	0.0191	0.0343	0.0217	n/a	n/a
II	49	3	0.0170	0.0522	0.0229	n/a	n/a
II	58	1	0.0002	0.0003	0.0003	n/a	n/a

BC	Team	Run	prec.	rec.	F_1	AP	FAP
II	58	2	0.0002	0.0003	0.0003	n/a	n/a
II	58	3	0.0003	0.0003	0.0003	n/a	n/a
II	60	1	0.0257	0.0709	0.0279	n/a	n/a
II	60	2	0.0053	0.0159	0.0061	n/a	n/a
II	60	3	0.0077	0.0172	0.0085	n/a	n/a
II.5	14	1	0.11728	0.13377	0.10619	0.02668	0.04264
II.5	14	2	0.08552	0.09484	0.08063	0.01623	0.02703
II.5	14	3	0.07262	0.18982	0.08491	0.03609	0.05065
II.5	14	4	0.09032	0.10918	0.09143	0.01994	0.03273
II.5	14	S08	0.02943	0.03005	0.02839	0.00179	0.00337
II.5	14	S25	0.06334	0.07711	0.05863	0.00629	0.01135
II.5	18	5	0.18540	0.15107	0.14191	0.03124	0.05121
II.5	22	1	0.00207	0.34900	0.00409	0.04569	0.00750
II.5	22	2	0.00049	0.43496	0.00097	0.03651	0.00189
II.5	22	3	0.00205	0.34490	0.00404	0.04578	0.00743
II.5	22	4	0.00051	0.44726	0.00102	0.04478	0.00200
II.5	22	5	0.00098	0.25257	0.00194	0.01920	0.00352
II.5	22	S05	0.00357	0.29937	0.00698	0.01452	0.00943
II.5	22	S10	0.00320	0.29774	0.00626	0.01111	0.00801
II.5	22	S11	0.00350	0.29390	0.00685	0.02113	0.01034
II.5	22	S12	0.00402	0.33305	0.00777	0.01072	0.00901
II.5	22	S13	0.00262	0.19393	0.00512	0.00496	0.00504
II.5	26	S16	0.00016	0.00615	0.00031	0.00000	0.00000
II.5	31	1	0.04918	0.02312	0.02693	0.00114	0.00218
II.5	31	2	0.03279	0.01765	0.01874	0.00058	0.00112
II.5	31	3	0.00271	0.06685	0.00483	0.00025	0.00048
II.5	31	4	0.00047	0.00546	0.00086	0.00000	0.00001
II.5	31	5	0.09426	0.05671	0.06092	0.00632	0.01146
II.5	31	S18	0.08115	0.03668	0.03604	0.00308	0.00568
II.5	32	1	0.05838	0.04781	0.04897	0.00243	0.00463
II.5	32	S07	0.05838	0.04781	0.04897	0.00380	0.00705
II.5	37	1	0.06803	0.23681	0.07706	0.02313	0.03558
II.5	37	2	0.06803	0.23681	0.07706	0.02101	0.03301
II.5	37	3	0.06803	0.23681	0.07706	0.02803	0.04110
II.5	37	4	0.07171	0.24501	0.07983	0.02770	0.04113
II.5	37	5	0.06793	0.23159	0.07106	0.02870	0.04089
II.5	42	S01	0.07335	0.10178	0.07619	0.00692	0.01270
II.5	42	S02	0.02063	0.20777	0.02874	0.00740	0.01177
II.5	42	S03	0.01780	0.16406	0.02843	0.00593	0.00981
II.5	42	S19	0.01721	0.06284	0.02338	0.00120	0.00229
II.5	42	S20	0.02424	0.29321	0.04164	0.01043	0.01668
II.5	51	1	0.00107	0.02596	0.00201	0.00033	0.00056
II.5	51	2	0.01073	0.07445	0.01569	0.00144	0.00264
II.5	51	3	0.00817	0.15988	0.01344	0.00356	0.00563

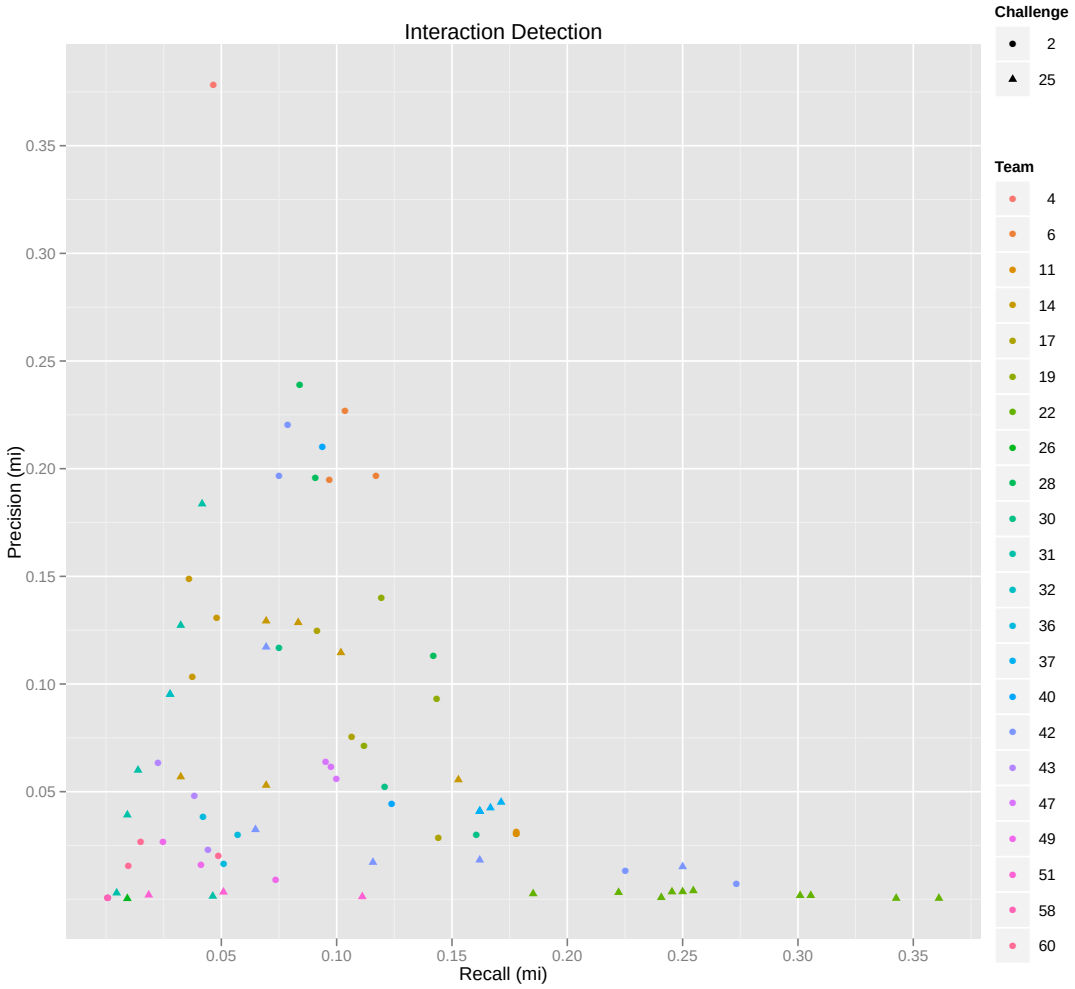

Figure A.3: Micro-averaged precision vs. recall plot of all interaction detection results. Triangles: BioCreative II.5 submissions, circles: BC II. Runs are color-coded per team.

A.1.4 Micro-averaged Method Extraction Results

BC	Team	Run	prec.	rec.	F_1	AP	FAP
II	14	1	0.3556	0.0732	0.1214	n/a	n/a
II	14	2	0.3281	0.0481	0.0838	n/a	n/a
II	14	3	0.3252	0.0606	0.1022	n/a	n/a
II	40	1	0.6679	0.2002	0.3081	n/a	n/a
II	40	2	0.4034	0.4302	0.4164	n/a	n/a
II	40	3	0.4965	0.4096	0.4489	n/a	n/a
III	65	1	0.08768	0.84820	0.15893	0.27588	0.20168
III	65	2	0.02448	1.00000	0.04779	0.23420	0.07938
III	65	3	0.09419	0.81784	0.16892	0.27727	0.20994
III	65	4	0.33483	0.42315	0.37385	0.13134	0.19438
III	65	5	0.02440	1.00000	0.04763	0.29016	0.08183
III	69	1	0.52065	0.55028	0.53506	0.34302	0.41804
III	69	2	0.54335	0.53510	0.53920	0.33824	0.41571
III	69	3	0.57359	0.50285	0.53589	0.32539	0.40492
III	69	4	0.59251	0.48008	0.53040	0.31711	0.39692
III	69	5	0.61333	0.43643	0.50998	0.29373	0.37276
III	70	1	0.48606	0.23150	0.31362	0.12948	0.18329
III	70	2	0.70000	0.11954	0.20421	0.08726	0.12227
III	70	3	0.80645	0.04744	0.08961	0.03796	0.05333
III	70	4	0.31220	0.36433	0.33625	0.15688	0.21394
III	70	5	0.32685	0.15939	0.21429	0.05734	0.09046
III	81	1	0.04540	0.66034	0.08496	0.07716	0.08087
III	81	2	0.08706	0.42125	0.14430	0.06239	0.08711
III	81	3	0.13514	0.28463	0.18326	0.04657	0.07427
III	81	4	0.13201	0.27704	0.17881	0.05601	0.08530
III	81	5	0.21350	0.22201	0.21767	0.05283	0.08502
III	88	1	0.28435	0.45161	0.34897	0.20244	0.25624
III	88	2	0.28172	0.45920	0.34921	0.20069	0.25489
III	89	1	0.52515	0.49526	0.50977	0.28202	0.36314
III	89	2	0.52016	0.48956	0.50440	0.28589	0.36494
III	89	3	0.50781	0.49336	0.50048	0.27238	0.35277
III	89	4	0.52495	0.49905	0.51167	0.29220	0.37198
III	89	5	0.52581	0.52182	0.52381	0.29969	0.38125
III	89	S4	0.52713	0.51613	0.52157	0.29926	0.38031
III	89	S5	0.52277	0.50095	0.51163	0.30046	0.37859
III	89	S6	0.52281	0.52182	0.52232	0.30044	0.38146
III	89	S7	0.49553	0.52562	0.51013	0.29303	0.37224
III	89	S8	0.51758	0.50285	0.51011	0.29766	0.37594
III	90	1	0.53333	0.47059	0.50000	0.26805	0.34900
III	90	2	0.52556	0.48767	0.50591	0.28386	0.36367
III	90	3	0.52300	0.58254	0.55117	0.35423	0.43128

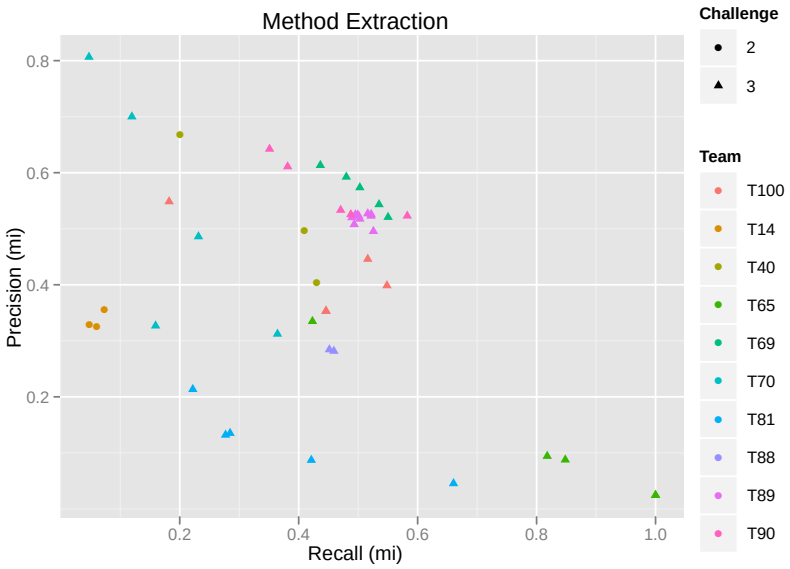

Figure A.4: Micro-averaged precision vs. recall plot of all 48 method extraction results. Triangles: BioCreative III submissions, circles: BC II. Runs are color-coded per team.

BC	Team	Run	prec.	rec.	F_1	AP	FAP
III	90	4	0.61094	0.38140	0.46963	0.25209	0.32808
III	90	5	0.64236	0.35104	0.45399	0.24270	0.31630
III	100	1	0.44590	0.51613	0.47845	0.26009	0.33699
III	100	2	0.39862	0.54839	0.46166	0.26982	0.34059
III	100	3	0.35285	0.44592	0.39396	0.14573	0.21276
III	100	4	0.35338	0.44592	0.39430	0.14583	0.21291
III	100	5	0.54857	0.18216	0.27350	0.11103	0.15794

A.2 PPI Article Estimation

This section describes how the PPI article estimate for 2009, quoted in the Introduction, Section 1.3.5, was taken. The combined number of articles indexed in Google Scholar for the year 2009 in the subject areas "Biology, Life Sciences, and Environmental Science" (Bio) and "Medicine, Pharmacology, and Veterinary Science" (Med) was (as of September 2010) 209,500 items: 122,000 in Bio + 145,000 in Med - 57,500 articles in both Bio and Med. This is a number similar to Highwire's 220,000 articles available for the same year. On the other hand, 20 to 30% of the articles of FEBS Letters, FEMBS Journal, EMBO Journal, or EMBO Reports contain protein interaction information according to the PPI curators[1], while many journals have no PPI data at all.

[1]Personal communication with Andrew Chatr-aryamontri.

Using a query, we *roughly* estimated the number of articles reporting PPI information in 2009. Terms describing methods to detect PPIs that are highly specific to this topic can be used to create the estimate. To ensure basic specificity of the results, the most distinct noun in the BioCreative II.5 corpus of PPI articles, "interaction", is combined with a method term using Boolean AND logic. The five most frequent PPI experimental method description terms used in the BioCreative II.5 corpus (and defined in the PSI-MI ontology) were calculated to be: coimmunoprecipitation, two hybrid, pull down, tandem affinity purification, and complementation assay. The final query is then constructed by requiring the noun interaction and any of these five terms:

> *interaction AND (coimmunoprecipitation OR 'two hybrid' OR 'pull down' OR 'tandem affinity purification' OR 'complementation assay')*

It is run against all articles in the two subject areas Bio and Med, but excluding patents, and only for the year 2009. Results are required to have "at least summaries" to be meaningful. This produces a result set of approximately 11,400 articles, or 5.4% of all indexed articles. Note that a similar query on PubMed would not re-produce this result, as the NLM data does not index the full text, only abstracts. As Google Scholar has full text access to most of the relevant journals, it is safe to assume that a conservative estimate of PPI-related (and thus, curation-relevant) articles for 2009 is at least 10,000 items, and unlikely to be larger than 10% of the published biomedical literature.

A.3 (Bio)NLP Terminology and Techniques

Extracting relationships (termed "events" in NLP) has a long standing history in BioNLP. In general, event extraction can be separated between approaches based on sentence co-occurrence of entity mentions, string/pattern matching methods, and linguistic methods based on parsing and chunking. Commonly, an NLP system is created from a mix of these approaches.

Most modern (Bio)NLP event extraction systems are based on statistical language models, and use (the frequency of) highly distinctive terms and context information as additional features. For example, in PPI extraction, the terms might include so-called *interaction verbs*, such as "interacts", "binds", "phosphorylates", etc., while context might gleaned be from the number of times two proteins are co-mentioned in the article's sentences, the position of the (potentially) PPI-containing sentence in the manuscript (title, abstract, heading, body, etc.), or other relevant entities co-mentioned in the vicinity of a protein.

The presence of a semantical relationship between terms (i.e., "entities") in a sentence is established by matching (parts of) the sentence to a (lexical) pattern or analyzing the syntactic structure of the sentence. In the simplest approach, if two entities are co-mentioned with a high frequency in several sentences, this co-occurrence is already a simple indicator to annotate a (likely) interaction.

These data are then used as features in supervised classifiers, commonly based on Naïve Bayes, logistic regression (maximum entropy) classifiers, kernel methods (e.g., SVMs), or Markov chains (e.g., CRFs). While unsupervised approaches do exist, they have been less frequently applied in BioNLP so far.

This development has lead to a rich jargon used in (Bio)NLP and makes it particularly difficult for outsiders to follow discussions related to natural language processing systems. A very brief introduction to the most important terms should help the reader follow the discourse of this work.

A.3.1 Phrases, Clauses, and Constituents

Phrases are complete blocks of a specific grammatical type (noun phrase, verb phrase, etc.) which can be treated as single units and can be replaced by a single word of the corresponding type where appropriate. For example, a noun phrase can be replaced by a single noun, even though the phrase spans several words or even nested phrases of other types.

Clauses are sequences of phrases and include at least one verb phrase. Just as phrases, they can be nested, too.

Phrases, clauses and sentences are all types of **constituents** – that is, hierarchically nested, grammatical units.

A.3.2 Tokens, Tokenization, and Tokenizers

A token essentially is a fancy way of referring to a single word or term. However, while the term "word" has a rather narrow definition and the word "term" an overly broad one, a token is simply a sequences of characters, be it letters, symbols or numbers found in the text defined by a deterministic process. A token can be a single word, a symbol, a sentence terminal (.!?:;), a numeric value, etc.. Depending on the **tokenization** strategy, words consisting of numbers, numerals and letters (as often found in protein and gene names) might be separated into multiple tokens for each character type. Some tokenization approaches might separate words on hyphens, dots, and parenthesis, while others might simply "tokenize" (i.e., separate the tokens) on white-spaces. The program that follows some algorithm for generating tokens is called a **tokenizer**.

A.3.3 Stemming and Lemmatizing

A **stemmer** removes the inflection from the stem of a word. For example, "historic", "history", and "histories" all can be normalized to the *stem* "histor". All variations of a word stem are known as its **surface variations**. Lemmatizing is the more elaborate process of not only removing inflections, but also finding the correct *lemma* or root form; For example, the lemma of "was" is "be", and of "better" it is "good". However, for "meeting" the lemma might be "meet" or "meeting" depending on whether the word is used as verb or noun, respectively. In other words, lemmatization provides a stricter regularization of words than stemming, but sometimes needs context to identify the correct lemma and always requires external knowledge representing mappings to the root of irregular lemmas. Both approaches are commonly used in NLP to normalize words for later processing steps.

A.3.4 Bag-of-words and Inverted Indices

An inverted index is a specialized form of a "bag-of-words". The "bag-of-words" is the unordered set of words or tokens found in a document. When supplemented with the the count how often each token was encountered (i.e., the words' frequencies), it is called an inverted index. Commonly, these sets and indices are created from the stemmed or lemmatized words to remove their surface variations. These inverted indices then are used as the features for a classifier. Therefore, already differences in the tokenization algorithm of two approaches can lead to different features provided to their classifiers.

A.3.5 N-grams

An n-gram is a sequence of n characters or tokens in a text or single sentence. Unigrams refer to a single items, bigram to two, a.s.f.. Commonly, token n-grams are used as features in a classifier, because the probability of these sequences is statistically more significant than single items (i.e., unigrams). For example, the difference in frequency of the trigram "yeast two-hybrid" found in PPI articles as opposed to its presence and frequency in other (non-PPI) articles will likely be significant. This stochastic property of text (just as with biological) sequences is often exploited to create more specific features than (unigram-based) bag-of-words would provide.

A.3.6 Part-of-speech Tagging

The part-of-speech (PoS) of a word is its linguistic category, such as noun, verb, adjective, etc.. As the PoS can be exploited as a feature in classification or used for syntactic analysis in the deep parsing and shallow chunking approaches described below, generating these categorical tags with a **PoS tagger** is a common procedure applied by NLP systems. PoS taggers are usually based on Markov (HMMs, CRFs) or maximum entropy (logistic regression) models and are trained on a collection of manually PoS-tagged sentences. PoS tags are lexical annotations.

In the process of Part-of-Speech (PoS) tagging, each word is assigned a grammatical concept (i.e., a part of speech) that describes its lexical class, such as verb, adjective, or noun. In PoS tagging, additionally the tense, number, and other syntactic categories are added, such as "verb, past tense", "adjective, superlative", or "noun, plural". The collection of all these syntactic categories is called a tag set. A commonly used PoS tagging scheme in BioNLP is the Penn Treebank tag set (Taylor et al. 2003). The PoS tags can be used as features for more advanced analysis techniques, for example, *chunking* to create segments (phrases, clauses) of semantically related words (e.g., verb or noun phrases), or *parsing* to establish the syntactic structure of a sentence. Commonly used PoS taggers are MedPost (Smith et al. 2004) or the GENIA PoS tagger (Tsuruoka et al. 2005), among others. As training data, so-called *corpora* are used, that are introduced in Section 1.1.3.

A.3.7 Parsing

Parsing refers to the technique of applying a mathematical formalism to produce a **syntax tree** (a directed acyclic graph) reflecting the syntactical relationships between words in the sentence of a language. The rules of a language can be formalized by a phrase structure (PS) grammar or a dependency grammar. For example, the PS grammar defining an arbitrary list of gene names in the English language might be defined as:

$$\begin{aligned} \mathit{Sentence} &\rightarrow \mathit{Name} \mid \mathit{List}\ \mathit{End} \\ \mathit{List} &\rightarrow \mathit{Name} \mid \mathit{Name}\ \texttt{,}\ \mathit{List} \\ \mathit{Name} &\rightarrow \texttt{A1BG} \mid \texttt{B2MR} \mid \texttt{CREB1} \mid ... \\ \texttt{,}\ \mathit{Name}\ \mathit{End} &\rightarrow \texttt{and}\ \mathit{Name} \end{aligned}$$

Each above line is called a *production rule*, the elements *Sentence*, *End*, *Name*, and *List* are *non-terminals*, while the commas, the "and", and the gene names are called *terminals*. A text is a valid language for a given grammar if, by following the production rules, all non-terminals can be resolved to terminal symbols. The above grammar defines a language that contains the single name "CREB1"[2], the sequence "A1BG, B2MR and CREB1"[3], or any combination of names including repetitions of the same name, as long as all names are separated by commas and the last name is separated by the conjunction "and". An example expression not part of this toy language is "A1BG, B2MR", because the non-terminal *End* introduced by the first production rule could not be resolved (The last rule ensures that an *End* can only be resolved if the *Name* is preceded by an "and" terminal.)

For Latin language grammars, the non-terminals would be phrases and the terminals the parts of speech (determiner, ajdective, noun, verb, etc.) of the words in the language, explaining the name of these grammars (*phrase* structure).

In addition to PS grammars, dependency-type grammars are the second main formalism used in NLP. In this case, the words or parts of speech (PoS) in the language themselves constitute the rules, without implicating the phrasal structure. Instead, each word or PoS can be said to directly *depend* on one or more others. While PS grammars are based on the work of Noam Chomsky, dependency grammars were first devisded by the work of Lucien Tesnière. Their syntax trees are therefore either hierarchical phrase structures (clustering phrases in the sentence according to the PS grammar rules) or dependency trees (commonly basing the tree's root at the predicate verb of the sentence). Beyond this fundamental binary separation, PS and dependency grammars are separated into many different sub-categories.

Parsers themselves are classified by the syntax tree they produce, the formal grammar, and the parsing algorithm they are based on. Beyond the grammar categories, multiple representations of syntax trees exist, and parsing techniques themselves are classified into different conceptual approaches (Grune & Jacobs 2008)[4].

[2] *Sentence* → *Name* → **CREB1**

[3] *Sentence* → *List* (→ *Name* (→ "**A1BG**"); "**,**"; *List* (→ *Name* (→ "**B2MR**"); "**,**"; *List* → *Name*)); *End* → "**and**"; *Name* → "**CREB1**"

[4] The first edition is freely available at `http://dickgrune.com/Books/PTAPG_1st_Edition/`

A.3.8 Parsers

The NLP technique of creating the syntax tree of a sentence in a language is referred to as (deep) **parsing**. Dependency trees create the simplest structures, and their parsers *tend* to be faster than parsers for other tree types. Commonly, for a parser to produce a syntax tree from a sentence, a *treebank* is used as training set; That is, a collection of sentences manually annotated with their particular syntax trees. Treebanks have been developed and optimized specifically for biomedical texts (e.g., the GENIA treebank). Most modern parsers "learn" a probabilistic model from the treebank data and then compute the similarity of input sentences to their model using kernel functions. It should be added that natural language parsers have been implemented using unsupervised approaches, too. However, (deep) parsers suffer from exponential (time) performance penalties.

A.3.9 Chunkers

To counter the performance issue of deep parsers, *shallow* parsers that only report *chunks* or boundaries of related terms in a sentence (e.g., a noun or verb phrase) have been developed (Aït-Mokhtar et al. 2002). For the obvious reason, these shallow techniques are often simply referred to as "**chunking**". Contrary to (deep) parsers, (shallow) chunkers do not report the syntactical relationships between the lexical constituents (i.e., chunks) in a sentence. While parsers can be used to determine the syntactical relationship of two constituents via the parse tree, the output of a chunker needs to be matched against lexical patterns to identify possible relations between mentioned entities, such as two protein mentions. Because these chunks are commonly determined by stochastic modeling (e.g., with MEMMs or CRFs) of a PoS-tagged token sequence and do not have to rely on an actual parsing algorithm, chunkers are (usually) faster than parsers.

In this work, the term "parser" (without the deep or shallow specifier) always refers to deep parsers.

A.3.10 Disambiguation

Disambiguation in NLP refers to the process of removing the ambiguity of a term. For example, "gene name disambiguation" is the procedure of resolving the correct gene identifier for a gene name that is assigned to genes across multiple species or in some cases even within a single species.

A.3.11 Coreference (Anaphora) Resolution

In linguistics, a co-reference occurs when an expression refers to another object in the same or some previous sentence. Typically, *coreferents* (a.k.a. *anaphora*) are pronouns (personal, possessive, demonstrative, indefinite, relative, or interrogative). The process of determining the correct object a pronoun is referring to is called coreference (or anaphora) resolution. Due to the inherent complexity by which co-references might be expressed (including forward references, a.k.a. cataphora), it is a topic of ongoing research.

Bibliography

Abi-Haidar, A., Kaur, J., Maguitman, A., Radivojac, P., Rechtsteiner, A., Verspoor, K., Wang, Z. & Rocha, L. M. (2008), 'Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks', *Genome Biol* **9 Suppl 2**, S11. 83

Agarwal, S., Liu, F. & Yu, H. (2011), 'Simple and efficient machine learning frameworks for identifying protein-protein interaction relevant articles and experimental methods used to study the interactions', *BMC Bioinformatics* **12**(Suppl 8), S10. 83

Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F. & Salakoski, T. (2008), 'All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning', *BMC Bioinformatics* **9 Suppl 11**, S2. 128

Aït-Mokhtar, S., Chanod, J. & Roux, C. (2002), 'Robustness beyond shallowness: incremental deep parsing', *Nat Lang Eng* **8**(3), 121–144. 167

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R. & Wang, X. (2008), Assisted curation: does text mining really help, *in* 'Pac Symp Biocomput', Vol. 13, Citeseer, pp. 556–7. 49, 145

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Tobin, R. & Wang, X. (2008), 'Automating curation using a natural language processing pipeline.', *Genome Biol* **9 Suppl 2**, S10. 83, 129, 131

Andrade, M. A. & Valencia, A. (1997), 'Automatic annotation for biological sequences by extraction of keywords from medline abstracts. development of a prototype system', *Proc Int Conf Intell Syst Mol Biol* **5**, 25–32. 23

Arighi, C., Lu, Z., Krallinger, M., Cohen, K., Wilbur, W., Valencia, A., Hirschman, L. & Wu, C. (2011), 'Overview of the biocreative iii workshop', *BMC Bioinformatics* **12**(Suppl 8), S1. 31, 39

Arighi, C. N., Roberts, P. M., Agarwal, S., Bhattacharya, S., Cesareni, G., Chatr-Aryamontri, A., Clematide, S., Gaudet, P., Giglio, M., Harrow, I., Huala, E., Krallinger, M., Leser, U., Li, D., Liu, F., Lu, Z., Maltais, L. J., Okazaki, N., Perfetto, L., Rinaldi, F., Saetre, R., Salgado, D., Srinivasan, P., Thomas, P. E., Toldo, L., Hirschman, L. & Wu, C. H. (2011), 'BioCreative III interactive task: an overview', *BMC Bioinformatics* **12**(Suppl 8), S4. 39, 46, 57

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000), 'Gene ontology: tool for the unification of biology. the gene ontology consortium', *Nat Genet* **25**(1), 25–9. 29, 30

Bada, M., Hunter, L., Eckert, M. & Palmer, M. (2010), An overview of the CRAFT concept annotation guidelines, *in* 'Proc Fourth Linguistic Annotation Workshop, LAW IV-10', Association for Computational Linguistics, pp. 207–211. 31, 131

Baker, D. (2001), 'Protein Structure Prediction and Structural Genomics', *Science* **294**(5540), 93–96. 18

Bard, J. B. L. & Rhee, S. Y. (2004), 'Ontologies in biology: design, applications and future challenges', *Nat Rev Genet* **5**(3), 213–22. 29

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C. & Apweiler, R. (2009), 'The goa database in 2009–an integrated gene ontology annotation resource', *Nucleic Acids Res* **37**(Database issue), D396–403. 40

Baumgartner, W. A., Lu, Z., Johnson, H. L., Caporaso, J. G., Paquette, J., Lindemann, A., White, E. K., Medvedeva, O., Cohen, K. B. & Hunter, L. (2008), 'Concept recognition for extracting protein interaction relations from biomedical text.', *Genome Biol* **9 Suppl 2**, S9. 83

Bell, J. (2004), 'Predicting disease using genomics.', *Nature* **429**(6990), 453–456. 17

Blake, J. & Bult, C. (2006), 'Beyond the data deluge: data integration and bio-ontologies.', *J Biomed Inform* **39**(3), 314–320. 29

Blaschke, C., Hirschman, L., Yeh, A. & Valencia, A. (2003), 'Critical assessment of information extraction systems in biology', *Comp Funct Genomics* **4**(6), 674–677. 36

Blaschke, C., Leon, E. A., Krallinger, M. & Valencia, A. (2005), 'Evaluation of BioCreAtIvE assessment of task 2', *BMC Bioinformatics* **6 Suppl 1**, S16. 39

Blaschke, C. & Valencia, A. (2001), 'The potential use of suiseki as a protein interaction discovery tool', *Genome Inform* **12**, 123–34. 18

Bodenreider, O. (2004), 'The unified medical language system (umls): integrating biomedical terminology', *Nucleic Acids Res* **32**(suppl 1), D267–D270. 22

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. (2007), 'Uniprotkb/swiss-prot', *Database* **2**, 3. 32

Brown, P. O. & Botstein, D. (1999), 'Exploring the new world of the genome with DNA microarrays.', *Nat Genet* **21**(1 Suppl), 33–37. 17

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K. & Wong, Y. W. (2005), 'Comparative experiments on learning information extractors for proteins and their interactions', *Artif Intell Med* **33**(2), 139–55. 31

Burns, G. A. P. C., Khan, A. M., Ghandeharizadeh, S., O'Neill, M. A. & Chen, Y.-S. (2003), 'Tools and approaches for the construction of knowledge models from the neuroscientific literature', *Neuroinformatics* **1**(1), 81–109. 111

Campbell, P. (2007), 'The database revolution', *Nature* **445**(7125), 229–230. 137

Cao, Y., Li, Z., Liu, F., Agarwal, S., Zhang, Q. & Yu, H. (2010), 'An ir-aided machine learning framework for the biocreative ii.5 challenge', *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 454–61. 83

Carroll, H. D., Kann, M. G., Sheetlin, S. L. & Spouge, J. L. (2010), 'Threshold average precision (tap-k): a measure of retrieval designed for bioinformatics', *Bioinformatics* **26**(14), 1708–13. 62, 115

Carroll, M. W. (2011), 'Why full open access matters', *PLoS Biol* **9**(11), e1001210. 28

Ceol, A., Chatr-Aryamontri, A., Licata, L. & Cesareni, G. (2008), 'Linking entries in protein interaction database to structured text: the febs letters experiment', *FEBS Lett* **582**(8), 1171–1177. 49

Chatr-aryamontri, A., Kerrien, S., Khadake, J., Orchard, S., Ceol, A., Licata, L., Castagnoli, L., Costa, S., Derow, C., Huntley, R., Aranda, B., Leroy, C., Thorneycroft, D., Apweiler, R., Cesareni, G. & Hermjakob, H. (2008), 'Mint and intact contribute to the second biocreative challenge: serving the text-mining community with high quality molecular interaction data', *Genome Biol* **9 Suppl 2**, S5. 18, 43, 46

Chatr-Aryamontri, A., Winter, A., Perfetto, L., Briganti, L., Licata, L., Iannuccelli, M., Castagnoli, L., Cesareni, G. & Tyers, M. (2011), 'Benchmarking of the 2010 BioCreative Challenge III text-mining competition by the BioGRID and MINT interaction databases', *BMC Bioinformatics* **12**(Suppl 8), S8. 43

Chaussabel, D. & Sher, A. (2002), 'Mining microarray expression data by literature profiling', *Genome Biol* **3**(10), RESEARCH0055. 19

Chen, H. & Sharp, B. M. (2004), 'Content-rich biological network constructed by mining pubmed abstracts', *BMC Bioinformatics* **5**, 147. 27

Chen, L., Liu, H. & Friedman, C. (2005), 'Gene name ambiguity of eukaryotic nomenclatures', *Bioinformatics* **21**(2), 248–56. 26

Chen, Y., Liu, F. & Manderick, B. (2010), 'Biolminer system: interaction normalization task and interaction pair task in the biocreative ii.5 challenge', *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 428–41. 83

Cochrane, G. R. & Galperin, M. Y. (2010), 'The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources', *Nucleic Acids Res* **38**(Database issue), D1–4. 17

Coghlan, A., Fiedler, T. J., McKay, S. J., Flicek, P., Harris, T. W., Blasiar, D., nGASP Consortium & Stein, L. D. (2008), 'ngasp–the nematode genome annotation assessment project', *BMC Bioinformatics* **9**, 549. 36

Cohen, A. M. & Hersh, W. R. (2005), 'A survey of current work in biomedical text mining', *Brief Bioinform* **6**(1), 57–71. 20

Corbett, P. & Copestake, A. (2008), 'Cascaded classifiers for confidence-based chemical named entity recognition', *BMC bioinformatics* **9**(Suppl 11), S4. 25

Couto, F. M., Silva, M. J., Lee, V., Dimmer, E., Camon, E., Apweiler, R., Kirsch, H. & Rebholz-Schuhmann, D. (2006), 'Goannotator: linking protein go annotations to evidence text', *J Biomed Discov Collab* **1**, 19. 43

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.-F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P. & Vidal, M. (2009), 'Literature-curated protein interaction datasets', *Nat Methods* **6**(1), 39–46. 113

Dai, H., Chang, Y., Tzong-Han Tsai, R. & Hsu, W. (2010), 'New challenges for biological text-mining in the next decade', *J Comp Sci Techn* **25**(1), 169–179. 42

Dai, H.-J., Lai, P.-T. & Tsai, R. T.-H. (2010), 'Multistage gene normalization and svm-based ranking for protein interactor extraction in full-text articles', *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 412–20. 83

Davis, J. & Goadrich, M. (2006), The relationship between precision-recall and roc curves, *in* 'Proc 23rd Int Conf Machine Learning, ICML-06', ACM, pp. 233–240. 115

Drmanac, R. (2011), 'The advent of personal genome sequencing', *Genet Med* **13**(3), 188–90. 17

Eaton, A. D. (2006), 'Hubmed: a web-based biomedical literature search interface', *Nucleic Acids Res* **34**(Web Server issue), W745–7. 23

Efthimiadis, E. (1996), 'Query expansion', *Annu Rev Inform Sci* **31**, 121–187. 138

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S.,

Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B. & de Jong, P. J. (2007), 'Identification and analysis of functional elements in 1% of the human genome by the encode pilot project', *Nature* **447**(7146), 799–816. 18

Erdman, A., Mangravite, L., Urban, T., Lagpacan, L., Castro, R., de la Cruz, M., Chan, W., Huang, C., Johns, S. & Kawamoto, M. (2006), 'The human organic anion transporter 3 (OAT3; SLC22A8): genetic variation and functional genomics', *Am J Physiol-Renal* **290**(4), F905–F912. 18

Evans, G. A. (2000), 'Designer science and the "omic" revolution.', *Nat Biotechnol* **18**(2), 127. 17

Falcon, S. & Gentleman, R. (2007), 'Using GOstats to test gene lists for GO term association.', *Bioinformatics* **23**(2), 257–258. 29

Fang, H., Murphy, K., Jin, Y., Kim, J. & White, P. (2006), Human gene name normalization using text matching with automatically extracted synonym dictionaries, *in* 'Proc Workshop Linking Natural Language Processing and Biology, BioNLP-06', Association for Computational Linguistics, pp. 41–48. 23

Fiehn, O. (2002), 'Metabolomics - the link between genotypes and phenotypes', *Plant Mol Bio* pp. 155–171. 17

Fields, S. & Song, O. (1989), 'A novel genetic system to detect protein-protein interactions', *Nature* **340**(6230), 245–6. 17

Fundel, K. & Zimmer, R. (2006), 'Gene and protein nomenclature in public databases', *BMC Bioinformatics* **7**, 372. 26

Gerstein, M., Seringhaus, M. & Fields, S. (2007), 'Structured digital abstract makes text mining easy.', *Nature* . 143

Grune, D. & Jacobs, C. (2008), *Parsing techniques: a practical guide*, Springer-Verlag New York Inc. 166

Hahn, U., Wermter, J., Blasczyk, R. & Horn, P. (2007), 'Text mining: powering the database revolution.', *Nature* . 137, 143

Hakenberg, J., Gerner, M., Haeussler, M., Solt, I., Plake, C., Schroeder, M., Gonzalez, G., Nenadic, G. & Bergman, C. M. (2011), 'The GNAT library for local and remote gene mention normalization.', *Bioinformatics* **27**(19), 2769–2771. 34

Hakenberg, J., Leaman, R., Vo, N. H., Jonnalagadda, S., Sullivan, R., Miller, C., Tari, L., Baral, C. & Gonzalez, G. (2010), 'Efficient extraction of protein-protein interactions from full-text articles', *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 481–94. 83, 131

Hakenberg, J., Plake, C., Leaman, R., Schroeder, M. & Gonzalez, G. (2008), 'Inter-species normalization of gene mentions with gnat', *Bioinformatics* **24**(16), i126–132. 27

Hakenberg, J., Plake, C., Royer, L., Strobelt, H., Leser, U. & Schroeder, M. (2008), 'Gene mention normalization and interaction extraction with context models and sentence motifs.', *Genome Biol* **9 Suppl 2**, S14. 83

Hamosh, A., Scott, A., Amberger, J., Bocchini, C. & McKusick, V. (2005), 'Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Res* **33**(suppl 1), D514–D517. 32

Hearst, M. A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M. A. & Ye, J. (2007), 'Biotext search engine: beyond abstract search', *Bioinformatics* **23**(16), 2196–7. 31

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A. et al. (2004), 'Intact: an open source molecular interaction database', *Nucleic Acids Res* **32**(suppl 1), D452–D455. 43

Hersh, W. & Bhupatiraju, R. (2003), Trec genomics track overview, *in* 'The Twelfth Text Retrieval Conference, TREC-03', pp. 14–23. 36

Hersh, W. & Voorhees, E. (2009), 'Trec genomics special issue overview', *Information Retrieval* **12**(1), 1–15. 36

Hirschman, L., Colosimo, M., Morgan, A. & Yeh, A. (2005), 'Overview of biocreative task 1b: normalized gene lists', *BMC Bioinformatics* **6 Suppl 1**, S11. 25, 33, 39, 40

Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. (2005), 'Overview of BioCreAtIvE: critical assessment of information extraction for biology.', *BMC Bioinformatics* **6 Suppl 1**(1471-2105 (Electronic)), S1. 31, 39

Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C. & Valencia, A. (2005), 'Text mining for metabolic pathways, signaling cascades, and protein networks', *Sci STKE* **2005**(283), pe21. 18

Hoffmann, R. & Valencia, A. (2005), 'Implementing the ihop concept for navigation of biomedical literature', *Bioinformatics* **21 Suppl 2**, ii252–8. 27

Hood, L., Heath, J. R., Phelps, M. E. & Lin, B. (2004), 'Systems biology and new technologies enable predictive and preventative medicine', *Science* **306**(5696), 640–3. 17

Huang, M., Ding, S., Wang, H. & Zhu, X. (2008), 'Mining physical protein-protein interactions from the literature', *Genome Biol* **9 Suppl 2**, S12. 83

Huang, M., Liu, J. & Zhu, X. (2011), 'Genetukit: a software for document-level gene normalization', *Bioinformatics* **27**(7), 1032–3. 27

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001), 'A comprehensive two-hybrid analysis to explore the yeast protein interactome', *Proc Natl Acad Sci U S A* **98**(8), 4569–74. 17

Jensen, L. J., Saric, J. & Bork, P. (2006), 'Literature mining for the biologist: from information retrieval to biological discovery', *Nat Rev Genet* **7**(2), 119–29. 20

Joachims, T. (1998), 'Text categorization with support vector machines: Learning with many relevant features', *Machine Learning: ECML-98* pp. 137–142. 24

Johnson, K. & Lin, S. (2001), 'Call to work together on microarray data analysis', *Nature* **411**(6840), 885. 36

Jones, K. (1972), 'A statistical interpretation of term specificity and its application in retrieval', *J Doc* **28**(1), 11–21. 24

Jose, H., Vadivukarasi, T. & Devakumar, J. (2007), 'Extraction of protein interaction data: a comparative analysis of methods in use', *EURASIP Journal on Bioinformatics and Systems Biology* p. 53096. 43

Joyce, A. R. & Palsson, B. Ø. (2006), 'The model organism as a system: integrating 'omics' data sets', *Nat Rev Mol Cell Bio* **7**(3), 198–210. 17

Kell, D. B. (2004), 'Metabolomics and systems biology: making sense of the soup', *Curr Opin Microbiol* **7**(3), 296–307. 17

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A., Vinod, N., Bader, G., Xenarios, I., Wojcik, J., Sherman, D. et al. (2007), 'Broadening the horizon–level 2.5 of the hupo-psi format for molecular interactions', *BMC Biology* **5**(1), 44. 43, 45, 48

Khoury, M. J., McCabe, L. L. & McCabe, E. R. B. (2003), 'Population screening in the age of genomic medicine', *N Engl J Med* **348**(1), 50–8. 17

Kiefer, J., Yin, H. H., Que, Q. Q. & Mousses, S. (2009), 'High-throughput sirna screening as a method of perturbation of biological systems and identification of targeted pathways coupled with compound screening', *Methods Mol Biol* **563**, 275–87. 19

Kiel, C., Vogt, A., Campagna, A., Chatr-Aryamontri, A., Swiatek-de Lange, M., Beer, M., Bolz, S., Mack, A. F., Kinkl, N., Cesareni, G., Serrano, L. & Ueffing, M. (2011), 'Structural and functional protein network analyses predict novel signaling functions for rhodopsin', *Mol Syst Biol* **7**, 551. 18, 48

Kim, J.-D., Ohta, T., Tateisi, Y. & Tsujii, J. (2003), 'Genia corpus–semantically annotated corpus for bio-textmining', *Bioinformatics* **19 Suppl 1**, i180–2. 31

Kim, J.-j. & Park, J. C. (2004), 'BioIE: retargetable information extraction and ontological annotation of biological interactions from the literature', *J Bioinf Comp Biol* **2**(3), 551–568. 29

Kim, J., Ohta, T., Pyysalo, S., Kano, Y. & Tsujii, J. (2009), Overview of bionlp'09 shared task on event extraction, *in* 'Proc Workshop Current Trends in Biomedical Natural Language Processing, BioNLP-09', Association for Computational Linguistics, pp. 1–9. 37

Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y. & Collier, N. (2004), Introduction to the bio-entity recognition task at jnlpba, *in* 'Proc Int Joint Workshop on Natural Language Processing in Biomedicine and its Applications, JNLPBA-04', Association for Computational Linguistics, pp. 70–75. 37

Kim, J., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N. & Tsujii, J. (2011), 'Overview of bionlp shared task 2011', *ACL HLT 2011* p. 1. 37, 112, 128

Kim, S. & Wilbur, W. (2011), 'Classifying protein-protein interaction articles using word and syntactic features', *BMC Bioinformatics* **12**(Suppl 8), S9. 83

Kitano, H. (2002), 'Systems Biology: A Brief Overview', *Science* **295**(5560), 1662–1664. 18

Koike, A., Niwa, Y. & Takagi, T. (2005), 'Automatic extraction of gene/protein biological functions from biomedical text', *Bioinformatics* **21**(7), 1227–1236. 29

Koike, A. & Takagi, T. (2004), 'Gene/protein/family name recognition in biomedical literature', *Proceedings of BioLINK 2004: Linking Biological Literature, Ontologies, and Databases* pp. 9–16. 33

Kolchinsky, A., Abi-Haidar, A., Kaur, J., Hamed, A. A. & Rocha, L. M. (2010), 'Classification of protein-protein interaction full-text documents using text and citation network features', *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 400–11. 83, 118

Krallinger, M., Izarzugaza, J. M. G., Rodriguez-Penagos, C. & Valencia, A. (2009), 'Extraction of human kinase mutations from literature, databases and genotyping studies', *BMC Bioinformatics* **10 Suppl 8**, S1. 18

Krallinger, M., Leitner, F., Rodriguez-Penagos, C. & Valencia, A. (2008), 'Overview of the protein-protein interaction annotation extraction task of biocreative ii', *Genome Biol* **9 Suppl 2**, S4. 39, 43, 55

Krallinger, M., Leitner, F. & Valencia, A. (2010), 'Analysis of biological processes and diseases using text mining approaches', *Methods Mol Biol* **593**, 341–82. 19

Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L. & Valencia, A. (2008), 'Evaluation of text-mining systems for biology: overview of the second biocreative community challenge', *Genome Biol* **9 Suppl 2**, S1. 31, 39

Krallinger, M., Tendulkar, A., Leitner, F., Chatr-aryamontri, A. & Valencia, A. (2010), 'The ppi affix dictionary (ppiad) and biomethod lexicon: importance of affixes and tags for recognition of entity mentions and experimental protein interactions', *BMC Bioinformatics* **11**(Suppl 5), O1. 34

Krallinger, M., Valencia, A. & Hirschman, L. (2008), 'Linking genes to literature: text mining, information extraction, and retrieval applications for biology', *Genome Biol* **9 Suppl 2**, S8. 20

Krallinger, M., Vazquez, M., Leitner, F., Salgado, D., Chatr-Aryamontri, A., Winter, A., Perfetto, L., Briganti, L., Licata, L., Iannuccelli, M., Castagnoli, L., Cesareni, G., Tyers, M., Schneider, G., Rinaldi, F., Leaman, R., Gonzalez, G., Matos, S., Kim, S., Wilbur, W., Rocha, L., Shatkay, H., Tendulkar, A. V., Agarwal, S., Liu, F., Wang, X., Rak, R., Noto, K., Elkan, C., Lu, Z., Dogan, R., Fontaine, J.-F., Andrade-Navarro, M. A. & Valencia, A. (2011), 'The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text', *BMC Bioinformatics* **12**(Suppl 8), S3. 39, 138

Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y. & Brunak, S. (2007), 'A human phenome-interactome network of protein complexes implicated in genetic disorders', *Nat Biotechnol* **25**(3), 309–16. 19

Lalonde, S., Ehrhardt, D. W., Loqué, D., Chen, J., Rhee, S. Y. & Frommer, W. B. (2008), 'Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations', *Plant J* **53**(4), 610–35. 20, 48

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A. & Huala, E. (2011), 'The arabidopsis information resource (tair): improved gene annotation and new tools', *Nucleic Acids Res* . 32

Lan, M. & Su, J. (2010), 'Empirical investigations into full-text protein interaction article categorization task (act) in the biocreative ii.5 challenge', *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 421–7. 83, 118

Leach, S. M., Tipney, H., Feng, W., Baumgartner, W. A., Kasliwal, P., Schuyler, R. P., Williams, T., Spritz, R. A. & Hunter, L. (2009), 'Biomedical discovery acceleration, with applications to craniofacial development', *PLoS Comput Biol* **5**(3), e1000215. 27

Leaman, R., Gonzalez, G. et al. (2008), Banner: An executable survey of advances in biomedical named entity recognition, *in* 'Pac Symp Biocomput', Vol. 13, pp. 652–663. 25

Leitner, F., Chatr-aryamontri, A., Mardis, S. A., Ceol, A., Krallinger, M., Licata, L., Hirschman, L., Cesareni, G. & Valencia, A. (2010), 'The febs letters/biocreative ii.5 experiment: making biological information accessible', *Nat Biotechnol* **28**(9), 897–9. 39, 46, 50

Leitner, F., Krallinger, M., Cesareni, G. & Valencia, A. (2010), 'The febs letters sda corpus: a collection of protein interaction articles with high quality annotations for the biocreative ii.5 online challenge and the text mining community', *FEBS Lett* **584**(19), 4129–30. 28, 31, 39

Leitner, F., Krallinger, M., Rodriguez-Penagos, C., Hakenberg, J., Plake, C., Kuo, C.-J., Hsu, C.-N., Tsai, R. T.-H., Hung, H.-C., Lau, W. W., Johnson, C. A., Saetre, R., Yoshida, K., Chen, Y. H., Kim, S., Shin, S.-Y., Zhang, B.-T., Baumgartner, Jr, W. A., Hunter, L., Haddow, B., Matthews, M., Wang, X., Ruch, P., Ehrler, F., Ozgür, A., Erkan, G., Radev, D. R., Krauthammer, M., Luong, T., Hoffmann, R., Sander, C. & Valencia, A. (2008), 'Introducing meta-services for biomedical information extraction', *Genome Biol* **9 Suppl 2**, S6. 39, 46, 47

Leitner, F., Mardis, S. A., Krallinger, M., Cesareni, G., Hirschman, L. A. & Valencia, A. (2010), 'An overview of biocreative ii.5', *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 385–99. 39, 140

Leitner, F. & Valencia, A. (2008), 'A text-mining perspective on the requirements for electronically annotated abstracts', *FEBS Lett* **582**(8), 1178–81. 136

Leser, U. & Hakenberg, J. (2005), 'What makes a gene name? named entity recognition in the biomedical literature', *Brief Bioinform* **6**(4), 357–69. 25

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., Castagnoli, L. & Cesareni, G. (2011), 'Mint, the molecular interaction database: 2012 update', *Nucleic Acids Res* . 43

Liu, H., Hu, Z., Zhang, J. & Wu, C. (2006), 'Biothesaurus: a web-based thesaurus of protein and gene names', *Bioinformatics* **22**(1), 103–105. 34

Loging, W., Harland, L. & Williams-Jones, B. (2007), 'High-throughput electronic biology: mining information for drug discovery', *Nat Rev Drug Discov* **6**(3), 220–30. 19

Lourenco, A., Conover, M., Wong, A., Nematzadeh, A., Pan, F., Shatkay, H. & Rocha, L. (2011), 'A linear classifier based on entity recognition tools and a statistical approach to method extraction in the protein-protein interaction literature', *BMC Bioinformatics* **12**(Suppl 8), S12. 83, 118

Lu, Z., Cohen, K. B. & Hunter, L. (2007), 'Generif quality assurance as summary revision', *Pac Symp Biocomput* pp. 269–80. 29

Lu, Z., Kao, H.-Y., Wei, C.-H., Huang, M., Liu, J., Kuo, C.-J., Hsu, C.-N., Tsai, R., Dai, H.-J., Okazaki, N., Cho, H.-C., Gerner, M., Solt, I., Agarwal, S., Liu, F., Vishnyakova, D., Ruch, P., Romacker, M., Rinaldi, F., Bhattacharya, S., Srinivasan, P., Liu, H., Torii, M., Matos, S., Campos, D., Verspoor, K., Livingston, K. M. & Wilbur, W. (2011), 'The gene normalization task in BioCreative III', *BMC Bioinformatics* **12**(Suppl 8), S2. 33, 39, 40

MacBeath, G. & Schreiber, S. L. (2000), 'Printing proteins as microarrays for high-throughput function determination.', *Science* **289**(5485), 1760–1763. 17

Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. (2011), 'Entrez gene: gene-centered information at ncbi', *Nucleic Acids Res* **39**(Database issue), D52–7. 32

Mandel, M. (2006), 'Integrated annotation of biomedical text: creating the pennbioie corpus', *Text Mining Ontologies and Natural Language Processing in Biomedicine, Manchester, UK* . 31

Manning, C., Schütze, H. & MITCogNet (1999), *Foundations of statistical natural language processing*, MIT Press. 25

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D. & Stolovitzky, G. (2010), 'Revealing strengths and weaknesses of methods for gene network inference', *Proc Natl Acad Sci U S A* **107**(14), 6286–91. 36

Martin-Sanchez, F., Iakovidis, I., Nørager, S., Maojo, V., de Groen, P., Van der Lei, J., Jones, T., Abraham-Fuchs, K., Apweiler, R., Babic, A., Baud, R., Breton, V., Cinquin, P., Doupi, P., Dugas, M., Eils, R., Engelbrecht, R., Ghazal, P., Jehenson, P., Kulikowski, C., Lampe, K., De Moor, G., Orphanoudakis, S., Rossing, N., Sarachan, B., Sousa, A., Spekowius, G., Thireos, G., Zahlmann, G., Zvárová, J., Hermosilla, I. & Vicente, F. J. (2004), 'Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care', *J Biomed Inform* **37**(1), 30–42. 18

McCray, A. T., Browne, A. C. & Bodenreider, O. (2002), 'The lexical properties of the gene ontology', *Proc AMIA Symp* pp. 504–8. 30

McQuilton, P., St Pierre, S. E., Thurmond, J. & the FlyBase Consortium (2011), 'Flybase 101 - the basics of navigating flybase', *Nucleic Acids Res* . 32

Metzker, M. L. (2010), 'Sequencing technologies - the next generation', *Nat Rev Genet* **11**(1), 31–46. 17

Miernyk, J. A. & Thelen, J. J. (2008), 'Biochemical approaches for discovering protein-protein interactions', *Plant J* **53**(4), 597–609. 20

Mooney, S. (2005), 'Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis.', *Briefings in Bioinformatics* **6**(1), 44–56. 18

Morgan, A. A., Lu, Z., Wang, X., Cohen, A. M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H.-h., Torres, R., Krauthammer, M., Lau, W. W., Liu, H., Hsu, C.-N., Schuemie, M., Cohen, K. B. & Hirschman, L. (2008), 'Overview of biocreative ii gene normalization', *Genome Biol* **9 Suppl 2**, S3. 33, 39, 40

Moult, J., Fidelis, K., Kryshtafovych, A. & Tramontano, A. (2011), 'Critical assessment of methods of protein structure prediction (casp)-round ix', *Proteins* **79 Suppl 10**, 1–5. 36

Müller, H.-M., Kenny, E. E. & Sternberg, P. W. (2004), 'Textpresso: an ontology-based information retrieval and extraction system for biological literature', *PLoS Biology* **2**(11), e309. 29

Mungall, C. J. (2004), 'Obol: integrating language and meaning in bio-ontologies', *Comp Funct Genomics* **5**(6-7), 509–20. 30

Nédellec, C. (2005), Learning language in logic-genic interaction extraction challenge, *in* 'Proc ICML-05, Learning Language in Logic Workshop, LLL-05', Citeseer, pp. 31–37. 31, 37

Neves, M. L., Carazo, J.-M. & Pascual-Montano, A. (2010), 'Moara: a java library for extracting and normalizing gene and protein mentions', *BMC Bioinformatics* **11**, 157. 27

Ongenaert, M., Van Neste, L., De Meyer, T., Menschaert, G., Bekaert, S. & Van Criekinge, W. (2008), 'Pubmeth: a cancer methylation database combining text-mining and expert annotation', *Nucleic Acids Res* **36**(Database issue), D842–6. 18

Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stümpflen, V., Ceol, A., Chatr-aryamontri, A., Armstrong, J., Woollard, P., Salama, J. J., Moore, S., Wojcik, J., Bader, G. D., Vidal, M., Cusick, M. E., Gerstein, M., Gavin, A.-C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., De Las Rivas, J., Prieto, C., Perreau, V. M., Hogue, C., Mewes, H.-W., Apweiler, R., Xenarios, I., Eisenberg, D., Cesareni, G. & Hermjakob, H. (2007), 'The minimum information required for reporting a molecular interaction experiment (mimix)', *Nat Biotechnol* **25**(8), 894–8. 45, 48, 143

Oti, M., Snel, B., Huynen, M. & Brunner, H. (2006), 'Predicting disease genes using protein-protein interactions', *J Med Genet* **43**(8), 691–698. 18, 48

Park, P. J. (2009), 'Chip-seq: advantages and challenges of a maturing technology', *Nat Rev Genet* **10**(10), 669–80. 17

Park, P. J., Pagano, M. & Bonetti, M. (2001), 'A nonparametric scoring algorithm for identifying informative genes from microarray data', *Pac Symp Biocomput* pp. 52–63. 19

Pedicini, M., Barrenäs, F., Clancy, T., Castiglione, F., Hovig, E., Kanduri, K., Santoni, D. & Benson, M. (2010), 'Combining network modeling and gene expression microarray analysis to explore the dynamics of th1 and th2 cell regulation', *PLoS Comput Biol* **6**(12), e1001032. 19

Perez-Iratxeta, C., Pérez, A. J., Bork, P. & Andrade, M. A. (2003), 'Update on xplormed: A web server for exploring scientific literature', *Nucleic Acids Res* **31**(13), 3866–8. 23

Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M. A. (2005), 'G2d: a tool for mining genes associated with disease', *BMC Genetics* **6**, 45. 18, 27

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K. B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G. N., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A. & Pandey, A. (2003), 'Development of human protein reference database as an initial platform for approaching systems biology in humans', *Genome Res* **13**(10), 2363–71. 18

Phan, I., Pilbout, S., Fleischmann, W. & Bairoch, A. (2003), 'Newt, a new taxonomy portal', *Nucleic Acids Res* **31**(13), 3822–3823. 32

Plake, C., Royer, L., Winnenburg, R., Hakenberg, J. & Schroeder, M. (2009), 'Gogene: gene annotation in the fast lane', *Nucleic Acids Res* **37**(Web Server issue), W300–4. 27

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J. & Salakoski, T. (2007), 'Bioinfer: a corpus for information extraction in the biomedical domain', *BMC Bioinformatics* **8**, 50. 31

Rabiner, L. & Juang, B. (1986), 'An introduction to hidden markov models', *IEEE/ASSP Magazine* **3**(1), 4–16. 25

Ramachandran, S. & Dash, D. (1999), 'The multitude of 'omics' and 'omes': Evolution of scientific terms in molecular biology in the new millennium', *Current Science Association* **77**, 846–847. 17

Ramakrishnan, C., Baumgartner Jr, W., Blake, J., Burns, G., Cohen, K., Drabkin, H., Eppig, J., Hovy, E., Hsu, C., Hunter, L., Ingulfsen, T., Onda, H., Pokkunuri, S., Riloff, E., Roeder, C. & Verspoor, K. (2010), 'Building the scientific knowledge mine (sciknowmine1): a community-driven framework for text mining tools in direct service to biocuration', *LREC - Language Resources and Evaluation Conference* . 111

Ratnaparkhi, A. (1996), A maximum entropy model for part-of-speech tagging, *in* 'Proc Empirical Methods in Natural Language Processing Conf, EMNLP-96', Vol. 1, pp. 133–142. 25

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C. O., Forrest, A. R. R., Gough, J., Grimmond, S., Han, J.-H., Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C. R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R. D., Tegnér, J., Lenhard, B., Teichmann, S. A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D. A., Ideker, T. & Hayashizaki, Y. (2010), 'An atlas of combinatorial transcriptional regulation in mouse and man', *Cell* **140**(5), 744–52. 20

Raychaudhuri, S., Chang, J., Sutphin, P. & Altman, R. (2002), 'Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature.', *Genome Res* **12**(1), 203–214. 29

Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M. & Stoehr, P. (2007), 'Ebimed–text crunching to gather facts for proteins from medline', *Bioinformatics* **23**(2), e237–44. 24

Rebholz-Schuhmann, D., Yepes, A. J. J., Van Mulligen, E. M., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E. & Hahn, U. (2010), 'Calbc silver standard corpus', *J Bioinform Comput Biol* **8**(1), 163–79. 31

Rebholz-Schuhmann, D., Yepes, A. J., Li, C., Kafkas, S., Lewin, I., Kang, N., Corbett, P., Milward, D., Buyko, E., Beisswanger, E., Hornbostel, K., Kouznetsov, A., Witte, R., Laurila, J. B., Baker, C. J., Kuo, C.-J., Clematide, S., Rinaldi, F., Farkas, R., Móra, G., Hara, K., Furlong, L. I., Rautschka, M., Neves, M. L., Pascual-Montano, A., Wei, Q., Collier, N., Chowdhury, M. F. M., Lavelli, A., Berlanga, R., Morante, R., Van Asch, V., Daelemans, W., Marina, J. L., van Mulligen, E., Kors, J. & Hahn, U. (2011), 'Assessment of ner solutions against the first and second calbc silver standard corpus', *J Biomed Semantics* **2 Suppl 5**, S11. 37, 110

Rinaldi, F., Kappeler, T., Kaljurand, K., Schneider, G., Klenner, M., Clematide, S., Hess, M., von Allmen, J.-M., Parisot, P., Romacker, M. & Vachon, T. (2008), 'Ontogene in biocreative ii', *Genome Biol* **9 Suppl 2**, S13. 83, 130

Rinaldi, F., Schneider, G., Kaljurand, K., Clematide, S., Vachon, T. & Romacker, M. (2010), 'Ontogene in biocreative ii.5', *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 472–80. 83

Rogers, F. et al. (1963), 'Medical subject headings.', *B Med Libr Assoc* **51**, 114. 22

Romano, P., Manniello, A., Aresu, O., Armento, M., Cesaro, M. & Parodi, B. (2009), 'Cell line data base: structure and recent improvements towards molecular authentication of human cell lines', *Nucleic Acids Res* **37**(Database issue), D925–32. 32

Rose, G. D. (1997), 'Protein folding and the paracelsus challenge', *Nat Struct Biol* **4**(7), 512–4. 36

Rubin, D. L., Thorn, C. F., Klein, T. E. & Altman, R. B. (2005), 'A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge', *J Am Med Inform Assoc* **12**(2), 121–9. 19

Saetre, R., Yoshida, K., Miwa, M., Matsuzaki, T., Kano, Y. & Tsujii, J. (2010), 'Extracting protein interactions from text with the unified akanere event extraction system', *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 442–53. 37, 83, 136

Sagae, K., Miyao, Y. & Tsujii, J. (2007), Hpsg parsing with shallow dependency constraints, *in* 'Proc Association for Computational Lingustics, ACL-07', Vol. 45, p. 624. 130

Sales, G., Coppe, A., Bisognin, A., Biasiolo, M., Bortoluzzi, S. & Romualdi, C. (2010), 'Magia, a web-based tool for mirna and genes integrated analysis', *Nucleic Acids Res* **38**(Web Server issue), W352–9. 19

Salton, G., Fox, E. & Wu, H. (1983), 'Extended boolean information retrieval', *Commun ACM* **26**(11), 1022–1036. 24

Salton, G., Wong, A. & Yang, C. (1975), 'A vector space model for automatic indexing', *Commun ACM* **18**(11), 613–620. 24

Salwinski, L., Licata, L., Winter, A., Thorneycroft, D., Khadake, J., Ceol, A., Aryamontri, A. C., Oughtred, R., Livstone, M., Boucher, L., Botstein, D., Dolinski, K., Berardini, T., Huala, E., Tyers, M., Eisenberg, D., Cesareni, G. & Hermjakob, H. (2009), 'Recurated protein interaction datasets', *Nat Methods* **6**(12), 860–1. 113

Schneider, G., Clematide, S. & Rinaldi, F. (2011), 'Detection of interaction articles and experimental methods in biomedical literature', *BMC Bioinformatics* **12**(Suppl 8), S13. 83, 118

Seal, R. L., Gordon, S. M., Lush, M. J., Wright, M. W. & Bruford, E. A. (2011), 'genenames.org: the hgnc resources in 2011', *Nucleic Acids Res* **39**(Database issue), D514–9. 26

Segura-Bedmar, I., Martínez, P. & de Pablo-Sánchez, C. (2011*a*), 'A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents', *BMC Bioinformatics* **12 Suppl 2**, S1. 31

Segura-Bedmar, I., Martínez, P. & de Pablo-Sánchez, C. (2011*b*), 'Using a shallow linguistic kernel for drug-drug interaction extraction', *J Biomed Inform* **44**(5), 789–804. 19

Sekimizu, Park & Tsujii (1998), 'Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts', *Genome Inform Ser Workshop Genome Inform* **9**, 62–71. 23

Seringhaus, M. R. & Gerstein, M. B. (2007), 'Publishing perishing? towards tomorrow's information architecture', *BMC Bioinformatics* **8**, 17. 137

Settles, B. (2005), 'ABNER: An open source tool for automatically tagging genes, proteins, and other entity names in text', *Bioinformatics* **21**(14), 3191–3192. 25

Sharoff, S. (2006), 'Open-source corpora: Using the net to fish for linguistic data', *Int J Corpus Ling* **11**(4), 435–462. 110

Simske, S. & Lin, X. (2004), Creating digital libraries: content generation and re-mastering, *in* 'First Int Workshop on Document Image Analysis for Libraries', IEEE, pp. 33–45. 28

Smalheiser, N. R., Torvik, V. I., Bischoff-Grethe, A., Burhans, L. B., Gabriel, M., Homayouni, R., Kashef, A., Martone, M. E., Perkins, G. A., Price, D. L., Talk, A. C. & West, R. (2006), 'Collaborative development of the arrowsmith two node search interface designed for laboratory investigators', *J Biomed Discov Collab* **1**, 8. 23

Smith, L., Rindflesch, T. & Wilbur, W. J. (2004), 'Medpost: a part-of-speech tagger for biomedical text', *Bioinformatics* **20**(14), 2320–1. 165

Smith, L., Tanabe, L. K., Ando, R. J. n., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C. M., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C. A., Povinelli, R. J., Vlachos, A., Baumgartner, Jr, W. A., Hunter, L., Carpenter, B., Tsai, R. T.-H., Dai, H.-J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., Maña-López, M., Mata, J. & Wilbur, W. J. (2008), 'Overview of biocreative ii gene mention recognition', *Genome Biol* **9 Suppl 2**, S2. 38, 39

Spasic, I., Ananiadou, S., McNaught, J. & Kumar, A. (2005), 'Text mining and ontologies in biomedicine: making sense of raw text', *Brief Bioinform* **6**(3), 239–51. 29

Sprinzak, E., Sattath, S. & Margalit, H. (2003), 'How reliable are experimental protein-protein interaction data?', *J Mol Biol* **327**(5), 919–23. 20

Stark, C., Breitkreutz, B.-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J. M., Winter, A., Dolinski, K. & Tyers, M. (2011), 'The biogrid interaction database: 2011 update', *Nucleic Acids Res* **39**(Database issue), D698–704. 18, 43

Starlinger, J., Leitner, F., Valencia, A. & Leser, U. (2009), Soa-based integration of text mining services, *in* 'Proc 2009 Congress on Services, SERVICES-09', IEEE, pp. 99–106. 136

Stewart, G., Crane, G. & Babeu, A. (2007), A new generation of textual corpora: mining corpora from very large collections, *in* 'Proc 7th ACM/IEEE-CS Joint Conf Digital Libraries', ACM, pp. 356–365. 28

Stumpf, M. P. H., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M. & Wiuf, C. (2008), 'Estimating the size of the human interactome', *Proc Natl Acad Sci U S A* **105**(19), 6959–64. 18

Sutton, C. & McCallum, A. (2007), *Introduction to statistical relational learning*, MIT Press, chapter An introduction to conditional random fields for relational learning, pp. 93–127. 25

Swanson, D. R. (1986), 'Fish oil, raynaud's syndrome, and undiscovered public knowledge', *Perspect Biol Med* **30**(1), 7–18. 22

Swets, J. (1969), 'Effectiveness of information retrieval methods', *American Documentation* **20**(1), 72–89. 114

Tanabe, L., Scherf, U., Smith, L. H., Lee, J. K., Hunter, L. & Weinstein, J. N. (1999), 'Medminer: an internet text-mining tool for biomedical information, with application to gene expression profiling', *Biotechniques* **27**(6), 1210–4, 1216–7. 19

Tanabe, L. & Wilbur, W. J. (2002), 'Tagging gene and protein names in biomedical text', *Bioinformatics* **18**(8), 1124–32. 31, 110

Tanabe, L., Xie, N., Thom, L. H., Matten, W. & Wilbur, W. J. (2005), 'Genetag: a tagged corpus for gene/protein named entity recognition', *BMC Bioinformatics* **6 Suppl 1**, S3. 31

Taylor, A., Marcus, M. & Santorini, B. (2003), The penn treebank: An overview, *in* A. Abeillé & N. Ide, eds, 'Treebanks', Vol. 20 of *Text, Speech and Language Technology*, Springer, pp. 1–22. 165

Templin, M. F., Stoll, D., Schrenk, M., Traub, P. C., Vöhringer, C. F. & Joos, T. O. (2002), 'Protein microarray technology', *Drug Discovery Today* **7**(15), 815–822. 17

Thompson, P., McNaught, J., Montemagni, S., Calzolari, N., Del Gratta, R., Lee, V., Marchi, S., Monachini, M., Pezik, P., Quochi, V., Rupp, C., Sasaki, Y., Venturi, G., Rebholz-Schuhmann, D. & Ananiadou, S. (2011), 'The biolexicon: a large-scale terminological resource for biomedical text mining', *BMC Bioinformatics* **12**, 397. 34

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J. & Leser, U. (2010), 'A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature', *PLoS Comput Biol* **6**, e1000837. 128

Tsuruoka, Y., McNaught, J., Tsujii, J. & Ananiadou, S. (2007), 'Learning string similarity measures for gene/protein name dictionary look-up using logistic regression', *Bioinformatics* **23**(20), 2768–74. 33

Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T., McNaught, J., Ananiadou, S. & Tsujii, J. (2005), 'Developing a robust part-of-speech tagger for biomedical text', *Advances in Informatics* pp. 382–392. 25, 165

Tsuruoka, Y., Tsujii, J. & Ananiadou, S. (2008), 'Facta: a text search engine for finding associated biomedical concepts', *Bioinformatics* **24**(21), 2559–60. 18

Tuason, O., Chen, L., Liu, H., Blake, J. A. & Friedman, C. (2004), 'Biological nomenclatures: a source of lexical knowledge and ambiguity', *Pac Symp Biocomput* pp. 238–49. 26

Valencia, A. (2002), 'Search and retrieve. large-scale data generation is becoming increasingly important in biological research. but how good are the tools to make sense of the data?', *EMBO Rep* **3**(5), 396–400. 137

Valencia, A. (2003), 'Meta, meta(n) and cyber servers', *Bioinformatics* **19**(7), 795. 125

Vanteru, B. C., Shaik, J. S. & Yeasin, M. (2008), 'Semantically linking and browsing PubMed abstracts with gene ontology', *BMC Genomics* **9 Suppl 1**, S10. 29

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden,

H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. (2001), 'The sequence of the human genome', *Science* **291**(5507), 1304–1351. 17

Verspoor, K., Cohen, K. B. & Hunter, L. (2009), 'The textual characteristics of traditional and open access scientific journals are similar', *BMC Bioinformatics* **10**, 183. 112

Verspoor, K., Roeder, C., Johnson, H. L., Cohen, K. B., Baumgartner, Jr, W. A. & Hunter, L. E. (2010), 'Exploring species-based strategies for gene normalization', *IEEE/ACM Trans Comput Biol Bioinform* **7**(3), 462–71. 83

Vidal, M., Brachmann, R. K., Fattaey, A., Harlow, E. & Boeke, J. D. (1996), 'Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and dna-protein interactions', *Proc Natl Acad Sci U S A* **93**(19), 10315–20. 17

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002), 'Comparative assessment of large-scale data sets of protein-protein interactions.', *Nature News* **417**(6887), 399–403. 17

Wang, X., Rak, R., Restificar, A., Nobata, C., Rupp, C., Batista-Navarro, R. T., Nawaz, R. & Ananiadou, S. (2011), 'Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature', *BMC Bioinformatics* **12**(Suppl 8), S11. 83

Wang, Y., Xiao, J., Suzek, T., Zhang, J., Wang, J. & Bryant, S. (2009), 'Pubchem: a public information system for analyzing bioactivities of small molecules', *Nucleic Acids Res* **37**(suppl 2), W623–W633. 32

Wang, Z., Kim, S., Quinney, S. K., Guo, Y., Hall, S. D., Rocha, L. M. & Li, L. (2009), 'Literature mining on pharmacokinetics numerical data: a feasibility study', *J Biomed Inform* **42**(4), 726–35. 19

Weston, A. D. & Hood, L. (2004), 'Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine', *J Proteome Res* **3**(2), 179–96. 17, 48

White, K. P. (2001), 'Functional genomics and the study of development, variation and evolution', *Nat Rev Genet* **2**(7), 528–537. 18

Wilbur, W. (1992), 'An information measure of retrieval performance', *Information Systems* **17**(4), 283–298. 114

Wilbur, W. & Coffee, L. (1994), 'The effectiveness of document neighboring in search enhancement', *Inform Process Manag* **30**(2), 253–266. 23

Wodak, S. J. (2007), 'From the mediterranean coast to the shores of lake ontario: Capri's premiere on the american continent', *Proteins* **69**(4), 697–8. 36

Yang, Y. & Pedersen, J. (1997), A comparative study on feature selection in text categorization, *in* 'Proc 14th Int Conf Machine Learning, ICML-97', Morgan Kaufmann Publishers, Inc., pp. 412–420. 118

Yeh, A., Morgan, A., Colosimo, M. & Hirschman, L. (2005), 'BioCreAtIvE task 1A: gene mention finding evaluation', *BMC Bioinformatics* **6 Suppl 1**, S2. 38, 39

Yeh, A. S., Hirschman, L. & Morgan, A. A. (2003), 'Evaluation of text data mining for database curation: lessons learned from the kdd challenge cup', *Bioinformatics* **19 Suppl 1**, i331–9. 35

Zheng, B., McLean, D. & Lu, X. (2006), 'Identifying biological concepts from a protein-related corpus with a probabilistic topic model.', *BMC Bioinformatics* **7**. 29

Zhu, H. & Snyder, M. (2003), 'Protein chip technology', *Curr Opin Chem Biol* **7**(1), 55–63. 17